

Blair's failure

Britain's research base is flourishing, and Tony Blair's last two governments can take much of the credit for it. But his third needs to focus on the troubled state of the universities.

When the Labour party came to power in Britain in 1997, it inherited a decaying science base staffed by disillusioned scientists. The fact that the country's main science lobbying group has recently changed its name from Save British Science to the Campaign for Science and Engineering (CaSE) says much about what the government has achieved since then. When the science minister David Sainsbury and his two opposition counterparts met for a press briefing last month, the spokesmen for both the Liberal Democrats and the Conservatives began by congratulating the Labour government for its record on science.

This lack of conflict was one reason why science played such a minor role in last week's UK elections. But although Labour has indeed performed impressively on this front, there are noticeable gaps in its record. Scientists and their organizations need to put pressure on Tony Blair's administration as it enters a third term in office.

Credit where it is due. Funding for science in universities and the research councils, the main source of UK grants, is up by more than 80% at £4.3 billion (US\$8.1 billion) annually since 1997 and is set to go on rising. Wisely, the critical problem of the nation's crumbling science infrastructure was tackled first. Sainsbury has consistently proved himself to be an enthusiastic science minister with well-tuned instincts for policies that are both sound and deliverable.

He has been greatly aided by the championship for science and associated wealth creation by the chancellor of the exchequer, Gordon Brown, who is widely expected to take over from Blair as prime minister within the next four years. The government has also employed a chief science adviser, David King, who has won the respect of researchers and the media and the ear of senior politicians.

Yet many UK scientists say it is as hard as ever to fund the basic research they want to do. Although more money is available, an increasing proportion of it is directed towards specific outcomes, and the freedom to fund responsively in some disciplines is succumbing to other priorities. CaSE calculates that the proportion of science funds controlled by central government has risen from 2% to 20% since 1997 — but that the funds available to the key funding agency for the great majority of physicists and chemists, the Engineering

and Physical Sciences Research Council, have risen by only 6%.

The government also faces a bigger challenge: how to reach the ambitious target, set last year, of raising total spending on research and development from 1.9% to 2.5% of gross domestic product by 2014. Everyone agrees that the gap must be made up predominantly by industry — everyone, that is, but industry itself, which, apart from the biomedical sector, by and large persists with a chronic lack of interest in research and is in long-term decline in several manufacturing sectors. Tax breaks and other measures announced last year will help, but many doubt whether the goal can be met. The government needs to abandon the target or explain how to meet it.

The continuing growth of biology-based industries will help, as will a focus on emerging and potentially research-intensive industries such as renewable energy. A drive towards applied research into low-carbon energy technologies could help bridge the gap between UK academics and the numerous small firms that work in this area.

But the scandal of the Blair government's record on science is to be found in the universities. There has been a haphazard response to the combination of declining interest among the young in science as a career — not unique to Britain — and misguided university funding schemes. The rise in the grant-funding science budget has not been accompanied by appropriate infrastructural and teaching support from higher-education funding councils. The latter have been too selective in favour of top-rated departments and have exposed the high costs of science departments to increasingly market-driven management. Behind it all lies a lack of joined-up government in addressing the supply and demand for future researchers. Meanwhile, the stream of department closures in the physical sciences is set to continue, to the increasing alarm of industry.

The balance between the freedom of universities to control their development and the nation's need to protect its knowledge and skills base is a delicate one. A critical weakness lies in the Department for Education and Skills, which is often criticized for a lack of responsiveness and excessive control. But it ultimately lies with Blair to ensure that the departments responsible for universities work coherently. In relation to science, this has been his most signal failure. ■

Proposals, please

You have one more month to submit proposals for ESOF2006, a fledgling but important forum for European science.

All credit to the small band of Europhile scientists behind the organization EuroScience. After a rush of blood to their heads, and with some skilled salesmanship, they overcame inertia, scepticism and indifference to launch in 2004 the first EuroScience Open Forum (ESOF).

The meeting attracted funding from foundations and the European Commission, and was supported by participants at the heart of European science, both individuals and institutions, including *Nature*. It attracted more than 1,800 participants from 67 countries, with 250 speakers, 50 sessions and 350 journalists. Feedback suggested that it achieved immediacy, relevance and comprehensibility,

in topics ranging from neuroscience to cosmology, and from research policy to science in schools.

The next meeting will be held in Munich, Germany, in July next year. It deserves to thrive but can only do so if Europe's scientists and citizens submit proposals. These should bring unusual collections of panellists together to address hot topics — leading perhaps to scenes like those last year, when sessions overflowed with people wanting to hear debates on climate change.

Those who, like *Nature*, wish to propose sessions can find the themes of ESOF2006 and submission forms at www.esof2006.org. The deadline for proposals is 15 June. ■



Nuclear fallout

Non-proliferation treaty in jeopardy as nations squabble
p132

Polio alert

Outbreak in Java delays programme of eradication
p133

**Clean cut**

Drop in pollution levels reverses global dimming
p135

**Going postal**

Famous scientists make it on to US stamps
p137

'Refusal to share' leaves agency struggling to monitor bird flu

Declan Butler, Paris

Tracking genetic changes in bird-flu viruses is vital for early warning of a human pandemic. But *Nature* has discovered that it is nearly eight months since the World Health Organization (WHO) last saw data on isolates from infected poultry in Asia. And from the dozens of patients who caught the deadly H5N1 strain this year, the WHO has managed to obtain just six samples.

Affected countries are failing, or refusing, to share their human samples with the WHO's influenza programme in Geneva. The UN Food and Agricultural Organization (FAO) set up a network of labs to collect animal samples last year, but it has not received any for months, and Michael Perdue, head of Animal Influenza Liaison at the WHO flu programme, complains that the FAO "hasn't been sharing" what it does have.

Such lack of cooperation is a key concern as anxiety about a possible pandemic increases. Human cases are beginning to appear in clusters, which suggests that people are transmitting the virus, older people are falling ill, and milder cases are being reported. Taken together, these trends suggest that the virus is becoming less virulent and more infectious — two characteristics typical of pandemic flu strains.

With so few samples to work on, it is impossible to judge how worried to be, says Klaus Stöhr, coordinator of the WHO's flu programme. "It's as if you hear a noise in your car engine, but you keep driving, not knowing whether it's serious."

Of the six human samples that the WHO has received from Vietnam, several contain a mutated version of H5N1. But that is not enough to indicate a broader change in the strain, says Perdue. It is also impossible for the agency to link this mutation of the virus to possible changes in how pathogenic and transmissible it is in humans. That would require molecular information on hundreds of viruses, and full clinical data on the cases from which they come. Such studies "aren't happening", says Stöhr.

Early signals that the virus is mutating might be picked up from viruses circulating in poultry. The FAO and the World Organisa-

Slim pickings: countries affected by avian flu are reluctant to release samples for outside analysis.

tion for Animal Health (OIE) should be collecting samples, but a recent FAO check reveals that the agency has not been receiving any. The WHO's flu programme was last given access to a sample in October 2004, so it has no idea how the virus is changing in birds.

Sensitive samples

Some countries don't have the resources to collect, conserve and securely transport samples, says Joseph Domenech, head of the Animal Health Service at FAO headquarters in Rome (see *Nature* 433, 102–104; 2005). "But things that should be happening are not," he adds. "Samples sometimes sit in labs," lacking authorization for export.

Countries are wary of sharing viruses with outside laboratories because they fear losing control over information, says one flu expert. "Authorities in Vietnam are very sensitive as to what they tell the people," he explains. "They don't want outside groups making pronouncements and these getting into the press without being vetted by the ministries of health and agriculture."

Scientists in countries with avian flu often want to work on virus samples first, he adds. They want to get credit for their work, he explains, and to use the data to develop their own vaccines.

One FAO consultant, who also asked not to be named, confirms there is a "time lag"

in sharing what samples there are with the WHO. But he argues that the FAO and OIE are in a difficult position. "Some countries have provided samples but stipulated that the information can't be shared with the wider community," he says.

Domenech argues instead that the FAO has no recent samples to share. There has been "complacency" at national levels, he admits, adding that the FAO has now instructed its regional networks to redouble their efforts to acquire isolates. And the FAO and OIE are drafting a standard 'material transfer' agreement to clarify the conditions of use of flu samples, and the intellectual-property rights of the countries that provide them.

Meanwhile, the WHO has begun soliciting poultry samples directly from affected countries. Stöhr, Perdue and other WHO officials flew to Manila in the Philippines last week to meet government health representatives from Vietnam, Cambodia and Laos. The talks included presentations on the mutated human strains.

The meeting heard that Vietnam has recently agreed to ship a large number of poultry samples direct to the WHO flu centre at the US Centers for Disease Control and Prevention in Atlanta, Georgia. And Perdue is hopeful that other countries will follow: "The presentations drove home the importance and urgency of sharing data." ■

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Nations spar over erosion of nuclear treaty

Geoff Brumfiel

The world's largest treaty to stop the spread of nuclear weapons is broken, but nobody can agree on how to fix it.

Representatives from more than 180 nations are meeting in New York over 2–27 May to discuss the Nuclear Non-Proliferation Treaty. But as *Nature* went to press in the meeting's second week, the atmosphere was bad tempered and participants had yet to agree an agenda.

The treaty requires nations with nuclear weapons to work towards disarmament, and those without to remain free of nuclear weapons in return for access to nuclear-power technology. Member states hold a meeting to review it once every five years.

Since it came into force in 1970, the treaty has been a bulwark against the proliferation of nuclear technologies. But in recent years, its authority has been increasingly challenged, and relations between member states have worsened. Activist groups have warned that agreeing a way forward at this month's meeting is essential if trust in the treaty is not to be eroded irreversibly (see *Nature* 433, 184; 2005).

Mohamed ElBaradei, director of the International Atomic Energy Agency, which polices the treaty, made the same warning to delegates when the conference opened. Without modification, he said, the treaty “will fade into irrelevance and leave us vulnerable and unprotected”.

Challenges to the treaty are coming from all sides (see ‘Problem States’ below). In 1998, India and Pakistan, who have not signed the treaty, tested nuclear devices. In 2003, North Korea announced its withdrawal from the treaty to pursue a nuclear weapons programme. Member states were left scrambling to work out how to respond, and lawyers are still arguing over whether North Korea's withdrawal was valid, as it was being investigated for suspicious activities when it decided to pull out.

Meanwhile, treaty members such as Brazil and Iran have been developing uranium-enrichment technologies that could allow them to produce a bomb in short order. In the case of Iran, many experts now agree that the goal of the programme probably is a weapon, not nuclear power as the country claims (see *Nature* 432, 432; 2004).

What's more, the benefits offered to non-weapons states for being part of the treaty are beginning to erode. Nuclear technology is now available on the black market, largely thanks to Abdul Qadeer Khan, the father of Pakistan's nuclear bomb. Last year Khan confessed to having headed an extensive network of scientists, engineers and businessmen who were selling nuclear secrets.

In addition, a rift is opening between

Mohamed ElBaradei warns an international meeting that its nuclear treaty may ‘fade into irrelevance’.

non-weapons states and the nuclear powers. Many of the former say that the five nuclear states — the United States, Britain, France, Russia and China — are failing to honour their obligations to disarm. Although these five nations have reduced the number of weapons in their stockpiles, critics point to their continued emphasis on the value of nuclear weapons as a deterrent. The United States in particular is developing new weapons, such as missile defence systems that could instigate a new arms race in space, and ground-penetrating ‘bunker busters’.

Most parties are agreed at least on which parts of the treaty need to be strengthened. Inspectors must be given the power to search more nuclear facilities and to conduct short-notice inspections. Withdrawal from the

treaty, which at present simply requires a three-month notice period, should be made more difficult. And many delegates want the weapons states to reaffirm their commitment to disarmament.

“There's a remarkable convergence about what the problems are,” says Rebecca Johnson, director of the London-based Acronym Institute for Disarmament Diplomacy, who is sitting in on the conference. “But there's very little agreement over how to deal with them.”

Delegates spent the first week of the meeting offering a bewildering range of opinions on what should be done. The US delegation wants to withhold enrichment technology, which is used to purify uranium fuel but also to make weapons-grade uranium, from countries such as Iran that have

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Problem states

The Nuclear Non-Proliferation Treaty has 187 signatories, but only a handful of these nations present real challenges to the treaty's future.

The ‘big five’ Under the treaty, the United States, Britain, China, Russia and France must all “pursue negotiations in good faith” towards disarmament. They say they are living up to their obligations, but other states say they must do more to limit their nuclear arsenals.

The wheelers and dealers Japan, the Netherlands, France, Russia and Germany all profit from the sale of nuclear technology abroad, and are loath to restrict where they peddle their wares. Their clients, such as Brazil and Iran, are equally set against new limits.

The outsiders Israel, Pakistan and India have developed nuclear weapons (pictured right), but are not signatories. Many of their neighbours signed the treaty in the hope that their regions would be nuclear-free — that hasn't happened.

North Korea It withdrew from the treaty in 2003, and may now be planning tests. Most experts agree that it was allowed to pull out too easily; many want to make withdrawal harder.

Iran The country is developing an advanced uranium-enrichment programme that until recently remained hidden from inspectors. It claims it is upholding its right to pursue the “use of nuclear energy for peaceful purposes” under the treaty. Others say its programme is a front for a bomb.

M. SEGAR/REUTERS

obstructed nuclear inspectors. Other suggestions included a moratorium on uranium enrichment, or allowing it only at plants under international control.

South Korea argued that withdrawal from the treaty should become contingent on approval by the United Nations Security Council, a rule that would have prevented North Korea from pulling out. And a host of countries, including Brazil and Egypt, called for substantial reductions in the nuclear arsenals of the five weapons states.

But each proposal raises its own issues among the other delegates. The United States is likely to resist any effort to curtail its nuclear arsenal. And placing restrictions on nuclear technology in certain states will meet resistance from members such as France and Germany, which profit from sales abroad.

The situation is so difficult that most observers doubt much will come of the conference. The treaty is unlikely to be amended for fear that the entire document may fall apart, says Michael Levi of the Brookings Institution, a non-partisan think-tank in Washington.

A more likely outcome could be a 'consensus document' on how to reinterpret the treaty. Such documents, which require the unanimous approval of all member states, have been created twice before: once in 1995 and again in 2000, in which a 13-step plan towards disarmament was proposed. A new one to strengthen the role of inspectors might stand a chance of success. Still, as Levi points out, "Iran would have to agree to it", which is unlikely. Other members might try to argue that the inspections of weapons states, which are currently voluntary, should be strengthened as well, but the 'big five' would be unlikely to approve.

Johnson is still hopeful that agreement will be reached, but says it won't be easy given the acrimonious tone of the conference. She calls on nations to make compromises — or risk nuclear weapons technology spreading out of control. "Quite a lot of states are protecting their own interests over and above international security," she fears. ■

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Waiting in line: Indonesia stepped up vaccinations after a Nigerian poliovirus paralysed four children.

Polio fight falters as Yemen and Java report fresh cases

Carina Dennis, Sydney

Polio is spreading to countries previously considered free of the disease, following a vaccine boycott in Nigeria in 2003. An outbreak in Yemen has sparked fears of an epidemic in the poorly immunized Middle Eastern nation. And the virus has now reached southeast Asia, with four cases confirmed in Indonesia last week.

Polio has not been seen in Yemen and Indonesia for a decade. Genetic analysis shows that the virus appearing in both areas is similar to the one that caused a 2003 outbreak in Nigeria. Polio vaccines were rejected in northern Nigerian states after Muslim clerics claimed they had been contaminated with HIV and contraceptives. Sixteen 'polio-free' countries have reported cases since 2003.

"The recent outbreaks can be traced back to that boycott," says Arun Thapa, an adviser on polio eradication in southeast Asia for the World Health Organization (WHO), based in New Delhi.

In April, the WHO confirmed 22 cases in Yemen, and health officials anticipate further infections because of low immunization rates among the nation's children. "We expect there will be many times this number," says Mohamed Wahdan, the WHO's Eastern Mediterranean polio adviser.

Middle Eastern nations have been on high alert for polio since December 2004 and many have already started preventative vaccination programmes. "Unfortunately,

the virus spread ahead of the campaign in Yemen," says Wahdan. "We hope to be ahead of the virus in other regions."

This month, Yemen will initiate a house-to-house campaign to vaccinate all its 3 million children under 5 years of age. Health officials are using a polio vaccine specifically targeted at the virus responsible for the outbreak, which they say provides greater immunity with fewer doses. "We believe it is the best tool in the face of an epidemic," says Wahdan.

In Indonesia, health officials are confident the virus can be restricted to the small villages in West Java where the four cases were reported. "The immunization level in Indonesia has been good, with 95% coverage of children," says Bardan Rana, the WHO's immunization officer in the country. But officials are taking no chances and are giving supplementary immunization to 5.2 million children under the age of five in West Java and surrounding provinces.

Although the recent outbreaks do not bode well for the eradication of polio, Thapa says he is more concerned about the situation in India. "It is the largest polio-endemic country and has been a source of wild poliovirus for other countries in the past."

Officials suspect the virus travelled to southeast Asia through Saudi Arabia, a popular destination for Indonesian workers and Muslim pilgrims. "As long as there is polio in the world and a lot of travel, we expect the virus to be imported," says Rana. ■

Competition boosts bid to find human genes

Alison Abbott

The human genome sequence has perplexed researchers from the moment the draft version was assembled in 2001. The problem: our genome seems to contain remarkably few protein-coding genes.

The current estimate is between 20,000 and 25,000 — not many more than far simpler organisms such as nematode worms. But pinning down the exact number has proved to be a laborious business, and efforts have so far made only limited progress. Bioinformaticians meeting in Cambridge, UK, last week were optimistic that they can reverse this trend, thanks to a competition called E-GASP.

Launched earlier this year, E-GASP challenged 18 teams from around the world to develop better gene-prediction software for the human genome.

The initiative has had the desired effect of improving the available gene-prediction software, says co-organizer Roderic Guigó, a bioinformatician at the Municipal Institute of Medical Research in Barcelona, Spain.

Proving that a particular stretch of DNA is a gene involves doing an experiment to show that it is transcribed to make an RNA copy that can then guide protein production. But to do this for the whole genome would be time-consuming and expensive. Software that predicts the likely position of genes can speed things up, but often has only limited accuracy.

E-GASP aimed to improve matters using test material taken from 44 regions of the human genome — about 1% of its total length. For 13 of the regions, researchers at ENCODE, a US initiative to analyse all of the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Researchers hope that advances in predictive software will speed the identification of genes.

functional elements in the human genome, painstakingly identified the position of all the genes by experiment.

This information was passed on to the 18 competing teams, who were then charged with predicting gene positions in the 31 remaining areas. At the same time, the ENCODE team completed its experimental analysis of the regions. Scientists gathered at the Wellcome Trust Sanger Institute on 6–7 May to hear the outcome.

“There was no absolute ‘right’ answer,” says Guigó. “Our annotation methods can only be described as ‘as-good-as-it-gets.’” So no overall winner was announced, although “a couple of the programs performed surprisingly well”, he adds.

Programs exploiting protein and transcription data provided the best predictions, but approaches involving comparisons with other genomes were also improved. Added together, the predictions put forward by the

competitors hit 70% of the genes identified by the ENCODE team almost perfectly.

Developing good prediction software is especially important for scientists working with species for which the genomes have been sequenced but little money is available for their analysis.

The new tools will also help guide the work of experimental scientists interested in human genes. The competitors’ predictions threw up hundreds of possible genes that weren’t identified in the lab experiments. ENCODE scientists in Barcelona and Geneva will select 200 of these for analysis in the next few months. “But based on our previous experience we do not expect more than 2% to be validated using our manual approach,” says Guigó.

He admits that other methods may turn up more genes. Researchers from the genomics company Affymetrix, based in Santa Clara, California, presented data to the Cambridge meeting from experiments using the latest generation of ‘microarrays’. These are made by chopping the genome up into thousands of bits of DNA and placing them, in order, on a grid. RNA will bind to the DNA it was produced from and therefore indicate any regions of the genome that are transcribed.

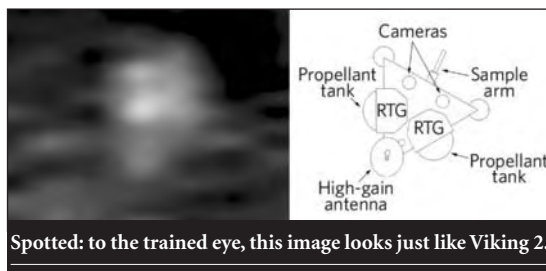
When the researchers washed RNA from a cell over the chip, 50% more regions on the grid bound to the RNA than there were known genes, suggesting that there is a lot more transcription going on than can be accounted for by genes identified so far. It isn’t known how much of this extra transcription represents new protein-coding genes, or whether some of the RNA molecules help to regulate existing genes. ■

Early martian visitors are caught on camera

Tony Reichhardt, Washington

Scientists operating the Mars Orbiter Camera have spotted a pair of long-lost spacecraft — the Viking Lander 2 and Mars Polar Lander — in pictures taken from martian orbit. They hope that the pictures of Mars Polar Lander can provide clues to how and why the spacecraft crashed.

The Mars Orbiter Camera has been circling the planet since 1997. Researchers at Malin Space Science Systems of San Diego, California, who built the camera with the California Institute of Technology, are fairly sure about the identification of Viking 2, which has been sitting on the surface since 1976. But confirmation for the polar lander, which is thought to have crashed when its braking rockets shut down prematurely



during its landing in 1999, will have to await sharper photos. Principal investigator Michael Malin hopes to begin taking those by late July, once frost on the martian surface has cleared up.

Malin has photographed other Mars landers from orbit, including Viking Lander 1, Mars Pathfinder and the two current Mars Exploration Rovers. But the panoramic

photos from Viking 2, which landed on a plain called Utopia Planitia, showed a flat and featureless terrain with few landmarks to help nail down where to look for it. Until now, the location was known only to within a few kilometres.

The picture believed to show the Mars Polar Lander was taken in 2000, although it was impossible to identify it at the time. Since then, photographs of the Mars rovers have given Malin a better idea of how parachutes and dark soil churned up by rocket blast would look from orbit.

So far, though, Malin hasn’t found anything in his pictures that looks like the small Beagle 2 lander lost in 2003. And until he has such a candidate, high-resolution searching of the surface would be like looking for a needle in a haystack. ■



Sun block: 'global dimming' caused by air pollution may have been masking the greenhouse effect.

Cleaner skies leave global warming forecasts uncertain

Quirin Schiermeier, Munich

Finally, some good news about the state of our planet. Studies of the amount of sunlight making it through the atmosphere suggest that our air is getting cleaner, thanks to reduced industrial emissions and the use of particulate filters.

But there's a nasty sting in the tail. Scientists are concerned that aerosols and dust in the air may have been shielding us from the worst of global warming. They don't know how extra solar radiation will affect future temperatures.

A downward trend in the amount of sunlight reaching the planet's surface, known as 'global dimming', has been noticed since measurements began in the late 1950s, but consensus that it was a global phenomenon was reached only last year (see *Nature* doi:10.1038/news040517-7; 2004). Many scientists have been reluctant to discuss the effect, fearing it would be used as an excuse to ignore the consequences of global warming.

They don't need to worry about that any more. Two studies, reported in *Science*, conclude that since 1990 the dimming has been replaced by brightening (M. Wild *et al.* *Science* **308**, 847–850; 2005 and R. T. Pinker, B. Zhang and E. G. Dutton *Science* **308**, 850–854; 2005). It has taken years to collect enough data for a statistically significant analysis, says Martin Wild, an atmospheric scientist at the Swiss Federal Institute of Technology in Zurich.

Wild and an international team of scientists analysed data from hundreds of ground stations around the world. They found that the amount of radiation reaching Earth's

surface fell by 4–6% between 1960 and 1990, but that the trend has since reversed nearly everywhere — although the total amount of radiation has not yet reached 1960 levels.

The result is backed up by a second study, led by Rachel Pinker from the University of Maryland, College Park, which infers a similar, albeit smaller, trend from satellite data.

"The good news is that the atmosphere has become cleaner and more transparent," says Andreas Macke, a meteorologist at the Leibniz Institute of Marine Sciences in Kiel, Germany. The collapse of communist economies in the late 1980s and the subsequent decrease in industrial pollutants released in the area was probably a major factor.

Wild and his team did detect continued dimming in some highly polluted areas, such as India, where vast clouds of smog from burning fossil fuels and wildfires darken the sky for long periods each year. But there was a brightening trend in China, despite the country's booming, fossil-fuel-intensive industry. "I am surprised," says Wild, adding that he can only speculate that the use of clean-air technologies in China may be more widespread and efficient than previously thought.

The question now is how the trend towards cleaner air will affect global temperatures. "It is clear that the greenhouse effect has been partly masked in the past by air pollution," says Macke.

Wild is investigating just how much was masked. He has yet to publish his results but he estimates that, until 1990, air pollution protected us from at least 50% of the warming that would have otherwise occurred. ■

Wanted: scientists to shape Europe's future research policy

Alison Abbott, Munich

Twenty people are set to become the most influential scientists in European basic research policy — at least for a while. They will form the governing council of the planned European Research Council (ERC), and they are due to be named next month.

The ERC will be the first pan-European research funding agency. Although part of the European Commission's next Framework programme for research (FP7), which begins late next year, it will be run by the academic community largely independently of the commission.

The first ERC governing council will be particularly powerful, as its remit will be to shape the broad programmes under which the European research community will apply for project funding. The commission has proposed that the ERC budget should average a hefty €1.5 billion (US\$1.9 billion) per year. But even if the European Parliament and the Council of Ministers agree to this generous funding, massive oversubscription is feared. So the council may try to limit demand by, for example, earmarking funds for certain sectors, such as young scientists. It will also set up evaluation and peer-review systems.

Members of the council are being selected by a panel of five academics chaired by Chris Patten, chancellor of Oxford and Newcastle universities. The panel was appointed by the European Commission in January and has since invited nominations from various European bodies involved in research, including national academies, research funding agencies, industry and universities. Despite requesting restraint, the panel has received well over 200 suggestions.

The final list will be designed to provide maximum credibility and authority, says the panel, and will be broadly representative of disciplines and types of research.

Gender and geography will also be taken into account, but panel members defend their commitment to idealism. "The ERC is about frontier research and excellence," says panel member and 1991 Nobel laureate Erwin Neher of the Max Planck Institute for Biophysical Chemistry in Göttingen. "The council needs to be credible in Europe, so balance is necessary — but there will be no question of geographical distribution of funds, or *juste retour*." ■

Los Alamos chief quits in wake of staff disaffection

San Diego The turbulent two-year reign of Los Alamos National Laboratory head Peter Nanos ended on 6 May with his resignation.

A former Navy admiral, Nanos was brought in after a series of embarrassing incidents at the lab involving classified material. But his aggressive management style prompted unrest among the lab's more than 10,000 employees, particularly when he accused researchers of operating in a "cowboy culture" (see *Nature* **433**, 447; 2005). One vocal insider website, "LANL: The Real Story", written by computer scientist Doug Richards, is said to have put particular pressure on Nanos.

Physicist Robert Kuckuck, a former senior administrator at the Lawrence Livermore National Laboratory in California, will take over from 16 May, with Nanos moving to the defence department.

Power plant tests reactions to Chernobyl-style disaster

Munich A nuclear power station in Romania was due this week to be the scene of a full-scale emergency operation for a nuclear

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Nuclear test: Romania's Cernavoda power station will simulate a full-scale emergency.

accident — but it's only a simulation.

Officials at the International Atomic Energy Agency (IAEA) will lead a team of seven international organizations on 11 May as they test the communication systems needed to deal with a Chernobyl-style accident. Plans for the initial reporting of an emergency and the health measures needed for those exposed to radioactive fallout will be assessed using a simulated disaster at the Cernavoda plant in eastern Romania.

The IAEA team will also track weather patterns to monitor where the fallout would land and which neighbouring countries to notify. Although the Chernobyl disaster occurred in what is now the Ukraine, much of the roughly 150,000 square kilometres of land that was contaminated lies to the north in Belarus.

Health agency logs on to beat fake drugs

London The World Health Organization (WHO) is harnessing the power of the Internet in its fight against counterfeit drugs — fake products that contain little, no or the wrong active ingredients.

Counterfeit drugs account for up to 10% of all medicines and are believed to kill thousands each year across the developing world (see *Nature* **434**, 132–136; 2005). They are smuggled across international borders, with many originating in China and India.

To combat this, the WHO wants health officials and pharmaceutical companies in the Asia-Pacific region to share information about fake drugs on a new website. The Rapid Alert System, launched last week at a meeting in Manila, the Philippines, will be useful if users are committed to sharing information, experts say. In the past, drug companies have proved reluctant to disclose information about counterfeiting.

♦ <http://218.111.249.28/ras>

Europe prepares to launch grand vision for space

Paris The European Space Agency (ESA) is this week set to put the finishing touches to its science programme for 2015–25. The final

version of the plan, known as Cosmic Vision, is so grand that ESA will need to forge new links with international partners, say agency advisers.

The proposals include sending a fleet of small spacecraft to Jupiter and its moons, putting a probe carrying clouds of ultracold atoms into orbit to test quantum gravity, and launching a space interferometer that can image planets outside the Solar System in the infrared.

The budget, to be decided in December, is expected to total €4 billion (US\$5 billion). Some big programmes will have to be done cooperatively, say ESA science advisers, but the success of the Cassini–Huygens mission to Saturn, a joint venture with NASA, and collaborations with Russia, China, Japan and India have paved the way, they add.

Who dares, wins a million for research

London The pot of cash for science prizes just got bigger. In an attempt to expand on the fields covered by the Nobel prizes, philanthropist Fred Kavli has created three million-dollar awards in nanotechnology, neuroscience and astrophysics.

Norwegian-born Kavli, who made his fortune in California selling sensors for aircraft, says he hopes the prizes will be

US gives stamp of approval to scientists

Washington Four stamps that bear the portraits of famous American scientists were released on 4 May by the US Postal Service.

The 37-cent stamps, which were created by artist Victor Stabin, include one that features the mathematician and physicist John von Neumann. Born in Hungary in 1903, von Neumann made significant contributions to game theory and quantum mechanics. He was also dubbed the ‘father of the computer’ for his theoretical studies of memory and logic circuits.

The three other stamps feature geneticist Barbara McClintock and physicists Josiah Willard Gibbs and

Richard Feynman. The US Postal Service has put out stamp series before on such themes as inventors or architects, but this is the first group of stamps featuring scientists.



more responsive to current research than the Nobels. “I think we’ll be more daring,” he says. The Nobels include awards for physics, chemistry and medicine, but are often bestowed years or even decades after the events that merited them.

The three new prizes, which will focus on basic research and be decided by a panel of international experts, will be awarded every two years from 2008. They will be presented in cooperation with the Norwegian Academy of Science and Letters and the Norwegian government.

Editorial note

A News story in the 21 October 2004 issue (*Nature* **431**, 889; 2004) reported that the US National Science Foundation (NSF) had been asked to investigate an allegation of scientific misconduct relating to the book *Born to Rebel: Birth Order, Family Dynamics, and Creative Lives* by Frank Sulloway, a visiting scholar at the Institute of Personality and Social Research at the University of California, Berkeley. The NSF Office of Inspector General has reviewed this allegation, found no evidence of data falsification, and closed the case.

Consenting adults? Not necessarily...

Companies and scientists in the West are keen to test their drugs in China, which is an important future market. But those running clinical trials need to be on their guard, says David Cyranoski.

In October 2003, China became the first country to move gene therapy into the medical mainstream, when it approved for general clinical use a treatment for head and neck squamous cancer. The feat was a symbol of the country's biomedical ambitions, and was lauded in the Chinese media.

But just a few months later, a story emerged that revealed a grimmer aspect of China's rapid development in biomedicine. In January, four HIV-positive farmers from Henan province, representing a further 15, sent a letter of complaint to the US National Institutes of Health and the Chinese Ministry of Health. They had participated in a clinical trial for a drug called VGV-1, and claimed that informed-consent procedures had been neglected.

The farmers are among hundreds of thousands of people in Henan infected with HIV by unhygienic commercial blood-donation practices during the 1990s. They say that they were not told about the risk of side effects from VGV-1, and are now consulting lawyers about a possible claim for compensation against Beijing's Ditan Hospital, which ran the trial — and possibly also against Viral Genetics of Azusa, California, the company that supplied the drug.

The episode is a sign that China's clinical research is jumping ahead of its system for ethical oversight. As *Nature* discovered on a recent visit, awareness of ethical regulations and informed-consent procedures is alarmingly low — even among researchers and medical professionals. Some institutional review boards (IRBs), which are supposed to enforce ethical standards, do little more than rubber-stamp the proposals that come before them. Xiaomei Zhai, a bioethicist at Peking Union Medical College, and a member of several IRBs, admits that many boards need to be better trained. "Otherwise we can't deal with the complicated problems that come up," she says.

There's a great deal of work to be done if China's biomedical researchers are to retain their people's trust as they stride into the clinic with experimental treatments. In the meantime, Western organizations

collaborating on clinical-research projects in China will need to take extreme care if they are to avoid getting their fingers burned like Viral Genetics. The company trusted its Chinese partners to deal with the regulatory and ethical aspects of the VGV-1 trial. Instead it has seen its name tarnished by the revelation that the project went ahead without approval from the State Food and Drug Administration (SFDA), which is supposed to check previous data to confirm that experimental drugs don't pose unacceptable hazards.

Land of opportunity

Despite the potential pitfalls, medical researchers are queuing up to work in China because of the plentiful clinical opportunities. Companies are also eyeing the long-term prospects of selling drugs into an emerging pharmaceuticals market. From rare genetic disorders, through obesity and other 'lifestyle' ailments, to a wide range of infectious diseases, China has one of the most complicated and debilitating disease profiles of any nation.

In theory, regulations put in place in 1999 mean that clinical trials in China must abide by Good Clinical Practice, a set of international standards that demand informed consent, approval of trial protocols by independent IRBs and the appropriate regulatory authorities, plus strict monitoring of ongoing trials. But in practice, enforcement of these standards is patchy. And unlike Western regulations, the Chinese rules contain no detailed guidelines on procedures for recruiting volunteers and ensuring informed consent. "We need details on the possibility of damage to subjects and about compensation," says Yiming Shao, a virologist at the National Center for AIDS/STD Prevention and Control in Beijing, who advises the Chinese government on public-health issues.



J. RUWITCH/REUTERS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

WU MING/KATZ PICTURES

A card recording the sale of blood in Henan (left) hides the price donors paid: many, like those pictured above, were infected with HIV

due to unhygienic collection practices.

What's more, a few diseases are deemed sufficient threats to public health that even some of those who are trying to tighten China's ethical framework are prepared to continue cutting corners. HIV, which could infect 10 million Chinese in the coming decade, heads this emergency list. "For AIDS,



Xiuping Li (left) and Jirong Pang say they weren't told of risks when they took part in an AIDS drug trial.

we have to accept that we have to compromise," says Zhai.

Such sentiments help explain how the VGV-1 trial fell into an ethical mire. VGV-1 consists of two proteins derived from calf thymus glands that, in lab tests, bind to an HIV protein called gp41. The idea is to prevent the virus from latching on to its target cells. A previous trial on ten patients in Mexico hinted that VGV-1 could restore the effectiveness of other AIDS drugs (J. J. Ayala-Gaytán *et al. HIV AIDS Rev.* 3, 8–13; 2004). The Chinese trial was designed to study the drug's effects in patients who had received no other antiretroviral therapy.

Recruitment drive

Trial participants tell alarming stories of how they were recruited. Xiuping Li, a farmer from Henan, claims she was told that the series of injections would put her in good health for 20 years without further treatment. "They didn't say that there were risks," she claims. Jirong Pang, another participant from Henan, says he suffered a high fever and red flaky rash all over his body and face. Other patients had more severe reactions, alleges Li, and were forced to pay for additional medicine to treat them.

Participants interviewed by *Nature* say they signed informed-consent forms that they could not understand and that doctors made no effort to explain. Their complaints go on: copies of the forms had to be paid for; expenses were not covered as agreed; participants weren't informed of the trial's results, despite asking. "That's common in China.

Most doctors don't answer patients' questions," says Fei Zang, a former accountant from Beijing who volunteered for the trial.

The patients' complaints were referred to the IRB of the National Center for AIDS/STD Prevention and Control. It recommended that, in future, doctors must make better efforts to explain trials to patients, and ruled that the 48 yuan (US\$5.80) charged for informed-consent forms should be returned and that participants should be paid expenses at a rate of 10 yuan per day. But overall, the panel concluded that there were no "serious problems" with the trial — much to the disappointment and anger of the participants.

Ruotao Wang, an epidemiologist at the Union School of Public Health in Beijing, who chairs the National Center for AIDS/STD Prevention and Control's IRB, explains that his remit didn't include the failure to get the drug approved for experimental use by the SFDA. "That's a serious problem," he says. The fact that the trial was approved by Ditan Hospital's IRB despite this omission underlines the inadequacy of the current framework for ethical oversight.

Zhai says that her IRBs would never authorize a trial without SFDA approval — but she understands how it might happen elsewhere. "A US company says it has medicine, and doctors say this is good news," Zhai says. On hearing this, inexperienced members of many Chinese IRBs may get carried away with enthusiasm. "Few have formal training in bioethics," says Zhai. "They are mostly scientists." Matters aren't

helped by the lack of any national IRB accreditation system.

Wang's IRB is one of the best regarded in China. Yet even its meetings are an eye-opening experience for anyone familiar with international standards — as *Nature* found out by attending a lively session in March. The committee evaluated eight proposals for new trials, three of which — including one collaboration with a US institution — detailed few or no procedures for informed consent. With a mix of laughter and censure, Wang's IRB interrogated the researchers. One scientist, realizing that her trials with children would require parental approval, thought that this should be enough. "She wanted to know whether approval from the children was also necessary," says Wang. His committee sent that proposal back to the drawing board.

Worryingly, such rulings may not have the desired effect of forcing researchers to adopt more rigorous ethical standards. "If we reject them, they often just go somewhere else," says Wang.

Paper trail

Procedures common in other countries, such as journal policies that stipulate ethical procedures, are also absent in China. "Most journals in China give no consideration to ethical matters," says one editor, who is adding notices of informed-consent procedures to his journal pages in an attempt to raise awareness among potential authors. "But if we followed international practice seriously, we would receive very few papers," he says.

Against this chaotic background, researchers from Western institutions can find themselves battling charges of impropriety even if they have satisfied international ethical requirements. David Ho, director of the Aaron Diamond AIDS Research Center in New York, who is conducting a trial of a combination of antiretroviral drugs in Yunnan province, has come under attack for not getting provincial ethical backing — even though well-regarded IRBs in Beijing and the United States approved the protocol. "Some people do not want us in China," claims Ho, who was born and raised in Taiwan. "They want to slow our programmes down because I am viewed as a foreigner."

Viral Genetics, meanwhile, is still counting the cost of its foray into the wilds of Chinese clinical research. "We appreciate the fact that perception of the trial may be tainted because of a bureaucratic mistake," says Alan Sheinwald, who handles public relations for the company. "But we believe that the underlying ethical, medical and scientific practices are up to any scrutiny that has been or will be brought to bear." That assertion may yet be tested in a court of law.

David Cyranoski is *Nature's* Asian-Pacific correspondent.

Rising star

As construction on the world's largest optical telescope nears completion in Spain, the country's astronomers are gearing up for an expanded role on the global stage. Mark Peplow follows the preparations for first light.



The vast silver dome of the Gran Telescopio Canarias (GTC) glitters as the sun beats down on the volcanic peak of Roque de los Muchachos. Here on La Palma, one of the Spanish Canary Islands, construction continues apace. When it finally opens for business at the end of this year, the GTC will take the title of the world's largest optical telescope — an impressive symbol of how far Spanish astronomy has come in a remarkably short time.

The telescope is also a testament to the lifelong vision of one man, a man who has made it his life's work to establish world-class observatories in the Canaries and put Spain at the forefront of astronomy research.

For ground-based telescopes, size and location are everything. When looking for dim and distant stars, the largest mirrors simply catch more photons, allowing astronomers to see farther and more clearly. With a mirror that is 10.4 metres wide, the GTC is unlikely to disappoint. And Los Muchachos is already home to 14 telescopes because, at 2,400 metres, its high altitude and cloudless weather reduce atmospheric distortion.

Tramping between the telescopes is like touring a futuristic martian outpost. Gleaming white buildings rise out from the red volcanic rocks, and a short stroll in the thin air quickly leaves me breathless. A perfectly still

layer of cloud hangs 1,000 metres below, covering the Atlantic Ocean in a white blanket broken only by neighbouring volcanic peaks.

In 1856, the first astronomer to visit these islands, Charles Piazzi Smyth, the Astronomer Royal for Scotland, declared Tenerife to be ideal for astronomy. But another 100 years passed before the rector of the University of La Laguna in Tenerife asked a new physics graduate called Francisco Sánchez to make detailed tests of a potential telescope site close to Mount Teide on the island.

"I quickly realized two things," recalls Sánchez. "That I was very attracted to astronomy; and that the site was incredibly good." From that point onwards, he devoted his life to making best use of this natural resource.

Leading light

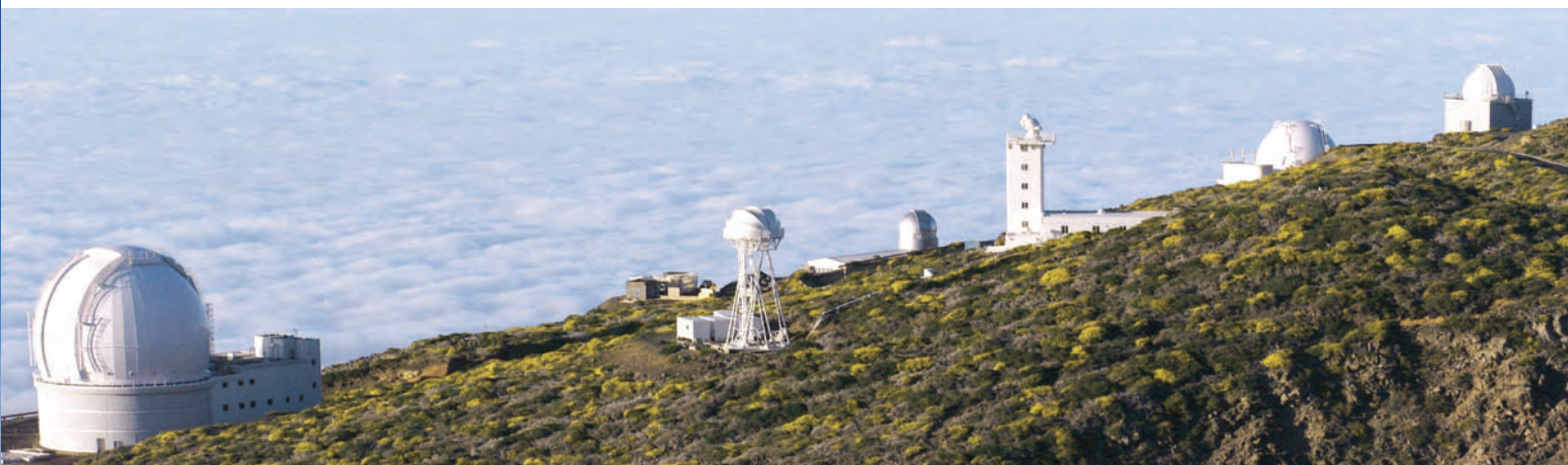
Sánchez is now director of the Canaries Institute of Astrophysics (IAC) in Tenerife, a post he has held since it was founded in 1975. When he began his investigations in the 1960s, there were four professional astronomers in Spain. Now there are more than 400, including 150 permanent research staff. Many of them credit Sánchez with single-handedly creating the vibrant research community from virtually nothing. "I just felt it would be a great shame if the islands were used only to tan people," Sánchez says.



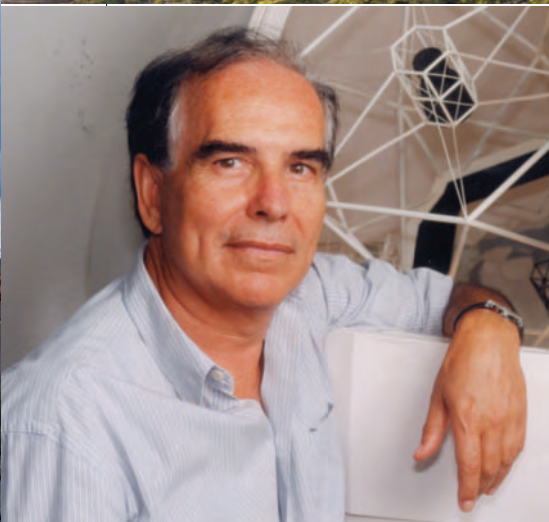
Jean-Claude Gérard and his wife take a tour of the Canaries' new telescope with Casiana Muñoz.

Although he now does little research, Sánchez commands great respect from his colleagues for the hard bargains he drives with politicians and international researchers. In person, Sánchez is genial, and speaks with the velvet-clad phrases of a consummate politician. "He's an autocrat, and he does get things wrong, but his right decisions wash away his sins," says John Beckman, a British astronomer who moved to the Canaries in 1984, and was the IAC's first research director.

Once Teide and Los Muchachos were established as prime locations for telescopes, Sánchez spent six years in negotiations to



Top of the world: the Gran Telescopio Canarias (opposite) is the latest addition to the instruments on Los Muchachos (above), largely thanks to the efforts of Spanish astronomer Francisco Sánchez (left).



IAC

ensure that whenever other nations wished to build observatories there, 20% of their observing time would be reserved for Spanish astronomers. Nineteen countries have telescopes in the Canaries, operated by more than 60 institutions worldwide. Part of the deal is that each country must also pay for several Spanish postdocs to work abroad each year. This was crucial in helping Spanish astronomy to come up to scratch so quickly, says Beckman.

Clear view

Sánchez says his ability to convince others stems from his own belief in the potential of the observatories. "I'm just a pathological optimist," he smiles.

His negotiating skills were also needed to convince the tourist industry, the Canaries' biggest money-spinner, to dim streetlights and neon signs. Legislation introduced in 1988 protects the observatories from light pollution, low-flying aircraft and air pollution. Public outreach activities have been essential to the observatories' progress, says Sánchez, who recently had a street named in his honour.

Strong local support for the IAC resulted in the regional government of the Canaries, together with the Spanish government, providing 90% of the €100 million (US\$128

million) funds for the GTC. Universities in Florida and Mexico have contributed the rest.

Many astronomers, even in Spain, doubted whether the GTC would be built, says José Miguel Rodríguez Espinosa, who leads the GTC project and is president of the Spanish Astronomical Society. The largest optical telescope built by Spain before the GTC was just 80 centimetres wide.

To convince funders that Los Muchachos was the perfect site for a giant telescope, IAC astrophysicist Casiana Muñoz used atmospheric physics to model the air turbulence above La Palma. She says that the site is as good as Hawaii, home to the world's largest optical instruments currently in operation — the two 10-metre Keck telescopes. The GTC is almost a direct copy of the Kecks, and will also use adaptive-optic techniques to modify its mirrors to compensate for turbulence in the air above. This should deliver images that rival space telescopes such as Hubble, Muñoz says proudly.

The GTC will house several special instruments to extract maximum information from the photons it collects. The first few are all Spanish-built. One of the most versatile optical instruments is OSIRIS, which takes spectroscopic measurements that can reveal the chemical composition of a light source. Such detailed measurements usually focus on individual, bright objects, but OSIRIS is designed to survey many dim objects over large parts of the sky. This means that it can be used to catalogue the ages and compositions of stars in the farthest reaches of our Galaxy.

Closer to home, OSIRIS will also focus on Kuiper-belt objects, frosty remnants from the formation of the Solar System that lie beyond the orbit of Neptune. Information about these primitive snowballs is limited to the largest and brightest objects, so OSIRIS should provide a more complete census.

Another instrument, EMIR, offers similar advantages at the near-infrared end of the optical spectrum. One of its first goals will be to peer deep inside the thick dust clouds that shroud stellar nurseries and capture the birth of a star on film. Because visible light is blocked by the clouds, astronomers hope that infrared surveys will allow them to see inside.

Broad scope

These instruments will not be operational until the end of 2006. Although engineering work on the GTC's 300-tonne superstructure is now complete, the project is two years behind schedule and the inside is still a building site. Wandering through the empty dome, I had a chance encounter with Jean-Claude Gérard, an astronomer from the University of Liège in Belgium, accompanied by his wife. They had taken a break from their holiday on La Palma to get a sneak preview of the building.

"It would be fantastic to be able to use it," Gérard says, admiring one of several hexagonal-mirror segments, each costing half-a-million dollars, that will soon be mounted on the telescope.

Gérard may well get his wish — the GTC has proved to be a

"Spain has been very good at using the telescopes other countries put on its land. Now it has the techniques needed to build its own facilities."

— Rafael Bachiller

powerful bargaining chip for Spanish astronomers who want to join the European Southern Observatory (ESO), a consortium of 11 European nations that share facilities such as the Very Large Telescope in Chile. On 28 April at a meeting in Madrid, Spanish and ESO negotiators confirmed that Spain would become a full ESO partner in 2006. The deal gives Spain a 25% reduction in its €65-million entry fee in return for a steady supply of data from the GTC to other ESO scientists.

"I think it's a good deal for Spain," says Rodríguez, a key player in the talks. Spanish astronomers will keep 90% of the observing time on the telescope, but some of this will be

IAC

earmarked for collaborative projects with ESO researchers.

Similar deals operate at the Keck telescopes, for example, where 85% of the time is reserved for researchers from Hawaiian and Californian universities or their research partners. Rodríguez is unapologetic. "If you own the telescope, your community uses the telescope," he says firmly.

But such comments reveal another side to the islands' success story, say several foreign researchers working there. "There is a degree of nationalism here that can be good and bad," says Mark Kidger, a British astronomer who moved to the Canaries with Beckman in the 1980s, and still has a contract position there. "It can be a great driving force, but there's also a reluctance in some areas to accept help from outside."

Closed shop

Because permanent research positions in Spain are regarded as civil-service jobs, Spanish applicants tend to be favoured. Those with foreign qualifications must undergo an expensive accreditation process, lasting several years, before they can apply to take the formal examinations required for an appointment. This bureaucracy will ultimately hinder Spanish astronomy's development, argues Kidger.

For example, the Armenian astronomer Garik Israelian, a rising star at the IAC, has been encouraged to take up Spanish citizenship to help with his career. "I have many Armenian colleagues who are professors in Britain and the United States who have had none of these problems," says Israelian. Still, he is keen to battle through the red tape. "I've had much better collaborations here because people are much more open to working together," he says.

Stories like Israelian's are not unusual. Nepotism is all too common, says Chris Benn, a British astronomer who works at the Isaac Newton Group of Telescopes on La Palma. Astronomy in Spain is "much more about having friends in the right places," he notes.

Spanish astronomers accept that there are problems, but insist the climate is changing. "Individualism has been the rule for many years in Spain, but now we're making large-scale collaborations with international partners," says Rafael Rebolo, one of the IAC's senior researchers. The IAC now requires students with new PhDs to spend at least two years working abroad before they return to the Canaries. "It gives a clear sign that these students must spend part of their career elsewhere," says Rebolo.

"Most Spanish astronomers agree that it would be good to open things up," says Johan Knapen, a Dutch astronomer at the University of Hertfordshire, UK, who worked with Beck-



Armenian Garik Israelian says he may need to take up Spanish citizenship to progress his career.

man in the Canaries. "But they don't want to if it means they won't get the jobs themselves."

Finding jobs for returning postdocs is one of the biggest problems facing Spanish astronomy, says Artemio Herrero Davó, head of research at the IAC. He is pushing for more government funding for junior positions. "Otherwise, young people will be discouraged and Spanish astronomy will lose its momentum," he says.

But most Spanish astronomers are positive about the future. In a global survey of astronomical research, Benn and his colleague Sebastian Sánchez (no

relation to Francisco) found that Spain now produces just under 5% of all astronomy papers, which is comparable to the output from Japan and the Netherlands (S. F. Sánchez and C. R. Benn *Astron. Nachr.* 325, 445–450; 2004). European

countries with a longer tradition in astronomy, such as Germany, France and Britain, produce about 10% each.

Treasure island

"The Spanish have made a decision to put a lot of money into astronomy, and it's growing faster than in any other country," says Benn. But much of that money has gone to the IAC, which some say has caused a bitter rivalry between the Canaries and astronomers on the mainland.

Rafael Bachiller, director of the Spanish National Astronomical Observatory (OAN) in Madrid, disputes this. "What happens in the Canaries is good for astronomy everywhere," he insists. "About 20 or 30 years ago, when optical and infrared astronomy were moving to the Canaries, the OAN decided to focus on radio astronomy," he says, adding

that the two institutions now have complementary rather than competing interests.

Bachiller is especially proud of the OAN's new 40-metre radio telescope at Yebes in central Spain, which he calls "an equivalent in radio waves to the GTC". The telescope was inaugurated in April, and will study the very first galaxies that evolved in the Universe.

Both the GTC and the 40-metre radio telescope are remarkable, says Bachiller, for being almost entirely Spanish-built. "Until now, we have been very good at using the telescopes that other countries put on our lands," says Bachiller, "but Spain has now developed the techniques needed to build its own facilities."

With the GTC close to completion and his legacy assured, Francisco Sánchez is expected to retire by the end of the decade. So will Spanish astronomy survive without his guiding hand? Many of his colleagues fall into a whisper as they explain that it's not possible to imagine astronomy without him. "I can't think of any single astronomer who could take over from Sánchez," says Israelian.

Others point out that he was an ideal leader during the 1970s, when funding was secured by charming a politician over a glass of wine. But now that research funding is far more structured, Sánchez's political skills are less essential for the IAC to flourish, they say.

Although Sánchez is cagey about naming his potential successors, Rebolo and Beckman are clearly front-runners. But for now, Sánchez is looking forward to returning to research after decades away from a telescope. "I'm going to become directly involved with research again on the GTC," he says, adding that he is particularly interested in studying planets outside our Solar System and understanding galactic structure. "I believe I've got the right to enjoy it," he laughs.

Mark Peplow is a reporter for news@nature.com.

Universities should foster neglected-disease work

Shifting the focus from patents and revenue to human welfare would speed progress.

Sir—Your Editorial “Wanted: social entrepreneurs” (*Nature* 434, 941; 2005) rightly points out the positive steps taken by not-for-profit pharmaceutical ventures to find cures for neglected diseases. But your gentle criticism of universities as impediments to these advances does not go far enough.

My experience is with a group known as Universities Allied for Essential Medicines (www.essentialmedicine.org), composed primarily of students seeking to hold universities to their avowed public mission in the arena of health-technology policy. We believe that universities’ reluctance to engage with non-traditional pharmaceutical partners stems in part from a myopic focus on taking out patents, executing licences and generating revenue. Success in technology transfer should be measured by its impact on human welfare, which requires an emphasis on innovation in neglected diseases and access to public-health goods.

The fact that neglected-disease drug ventures have to search and negotiate for molecules of interest reveals how upside-down the situation is. When patented

innovations have not yet been licensed to an external agency for further development, universities should allow other non-profit institutions to use them in research for neglected diseases, as a matter of policy. When innovations have been out-licensed, universities should include an exemption for research on neglected diseases in their licensing agreements. In either case, the university should forgo royalty payments on products sold in developing countries.

These exemptions can be constructed in such a way that they create a ‘dual-market opportunity’: any products developed could require cross-licensing (agreements between the beneficiary of the neglected-disease exemption and the original licensee) for sale in high-income countries, while being sold in poor countries without further licensing or payment of royalties.

Another point is that the criteria for academic promotion reinforce the difficulty of translating basic research into end-products. In addition to publications and grants, universities should consider a candidate’s work in finding treatments for

neglected diseases. They could reward participation in preclinical development projects, particularly open-source initiatives pooling research resources to speed commercialization, such as Tropical Disease Initiative (www.tropicaldisease.org) and Biological Innovation for Open Society (www.bios.net).

Of course, we are a long way from having neglected-disease research free of such hurdles. In the meantime, universities should look to their peers who are leading the charge in overturning the status quo. Yale University, the University of Washington, the University of California, Berkeley, the University of California, Santa Barbara and the University of Nebraska have all struck deals with non-traditional pharmaceutical ventures transferring intellectual-property rights to further neglected-disease research.

One hopes they are out in front of a much broader trend.

Dave A. Chokshi

Trinity College, University of Oxford,
Oxford OX1 3BH, UK

Two-stage drug approval would reduce the risks

Sir—Your timely News Feature “The safety catch” (*Nature* 434, 554–556; 2005) points out the need for better post-marketing safety surveillance of new drugs by the US Food and Drug Administration. At present, after a drug is approved on the basis of trial results, its effects in widespread use are difficult to assess and authorities have little power to control its promotion. Let us consider the consequences of a two-stage approval process for new drugs.

During the period of a few years between initial and final approval, the drug would not be promoted to doctors or (as is legal in the United States) to patients; it would be used only by patients whose conditions had not responded to existing drugs. In return, the clock would be stopped on patent expiration, which would compensate the manufacturer for the delay in promoting the drug more widely.

The public would benefit from the gathering of necessary safety data; the risk would be taken by a limited number of patients, in return for the chance of a more helpful medicine. Both manufacturer and public would benefit from the saved cost of not promoting new drugs that fail the post-marketing safety surveillance.

Incidentally, a similar but more

prolonged postponing of patent expiration on new antibiotics would help to delay bacterial resistance to them, because they would not be used initially unless the previously available agents were ineffective. This would also make the economics of limited use of new antibiotics much more fair to the manufacturers.

John A. Frantz

Monroe Clinic, 515 22nd Avenue, Monroe,
Wisconsin 53566, USA

Seeing clearly is not necessarily believing

Sir—I was interested to see your News Feature “CSI: cell biology” (*Nature* 434, 952–953; 2005) on digital photography and image manipulation in cell biology.

Photoshop-based enhancement of images raises questions of proper conduct, for which journal guidelines are necessary.

In my field, the problem (and it is here a problem, not an issue of misconduct) is much greater in electron than in light microscopy.

An example of good practice is the cover of the 2 December issue of *Nature*. This enhanced image is taken from an Article by A. Fotin and colleagues on clathrin lattices (*Nature* 432, 573–579; 2004). The authors are to be congratulated

for their unusually open disclosure of the difference between the original and the published image.

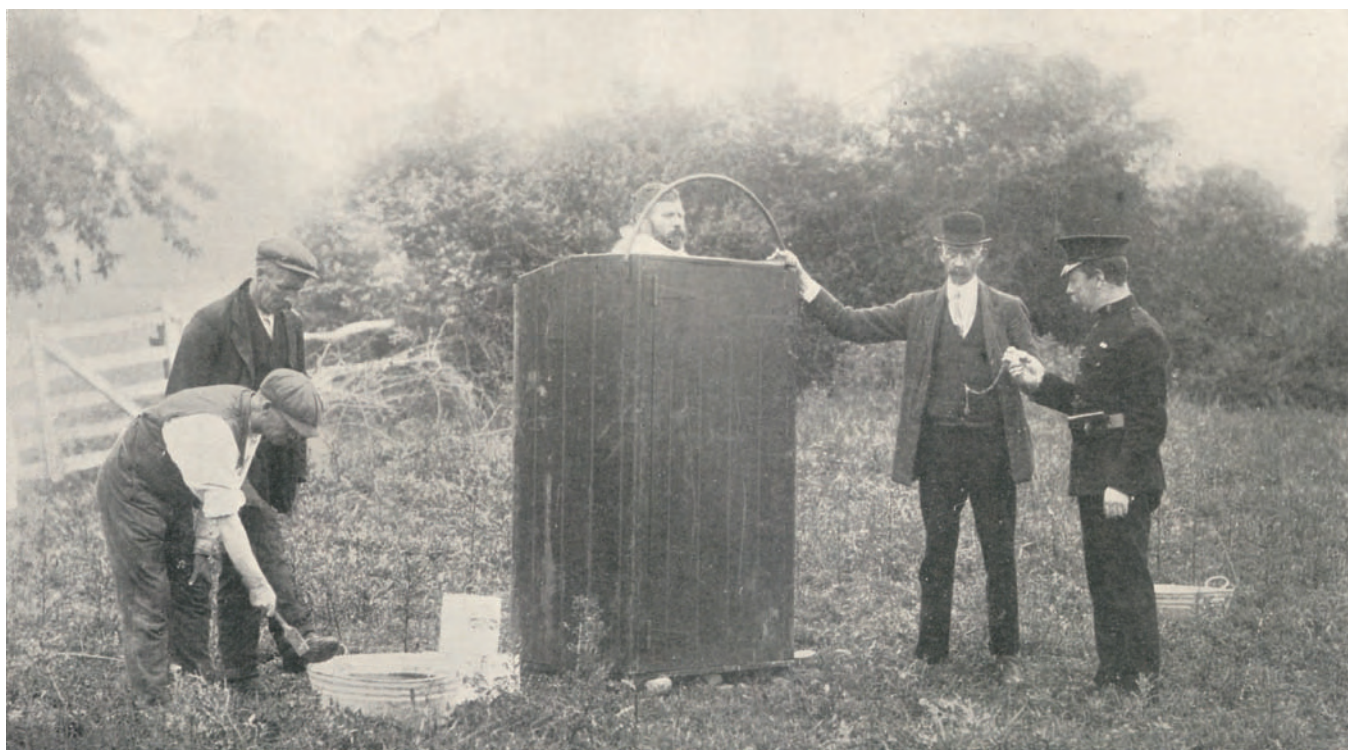
The Methods section of this Article makes a clear distinction between the data collected by the cryo-microscope and the pictures in the article. The authors list all the image-processing programs used in preparation of the illustrations: IMAGIC, CTFILT, FREALIGN, O, EMAN, Chimera, MAVE, LSQ_EXPLICIT, MAMA and MODELLER. These are mostly ‘off the shelf’ programs that produce symmetry enhancements, density averaging and many of the same effects as Photoshop.

In my experience, unless the scientist/postdoc/technician knows a great deal more about the guts of these programs than most, they are performing ‘black-box’ image enhancements that they do not control to any significant degree.

The full disclosure by Fotin and colleagues is remarkable for being so rare. The scarcity of such imaging disclosures elsewhere in the published record shows us just how far we have come towards inverting the purpose of scientific images. Where we used to have “seeing is believing”, we now have the possibility of “believing is seeing”, courtesy of our image-processing and enhancement software.

Mott T. Greene

Honors Program, CMB 1061, University of Puget Sound, Tacoma, Washington 98416-1061, USA



Control measure: disinfectant sprays have been used for a century to stop the spread of foot-and-mouth disease.

Death on the farm

Is the control of foot-and-mouth worse than the disease?

A Manufactured Plague: The History of Foot-and-Mouth Disease in Britain

by Abigail Woods

Earthscan: 2004. 224 pp. £19.99, \$35

Matthew Baylis

The 2001 epidemic of foot-and-mouth disease (FMD) in Britain will long be remembered by the farming community. But why? Because of the terrible suffering and high death rate of animals unfortunate enough to be infected? Or is it little to do with the disease itself and more to do with the reaction to it?

In this excellent book, Abigail Woods presents a compelling case that during FMD epidemics in Britain over the past century or so, it was the action taken by the agencies charged with controlling the disease that caused suffering in the farming community, rather than FMD itself. The disease is not, Woods argues, as fearsome as it is widely perceived to be. Foot-and-mouth disease is not a 'natural plague' of animals like, for example, rinderpest, which can spread rapidly and has, in the past, killed millions during pandemics. Rather, it is a manufactured plague — the suffering associated with it is largely of our own making.

How did this come about? Woods provides a fascinating account of the history of

FMD in Britain, from its first discovery in 1839 to its most recent appearance in 2001. She shows how the disease was transformed from being a minor affliction to one of the most feared of all livestock diseases.

A key event in this transformation was the 1865–67 rinderpest epidemic in Britain, which killed one out of every 15 cattle. Traditional control measures failed and the disease was finally checked by slaughtering all the infected animals and their contacts, and placing restrictions on cattle movements. These control measures were found to reduce the spread of FMD as well. However, at the time FMD was considered to be a mild disease and, accordingly, the UK government's Contagious Diseases (Animals) Act, 1869, laid out more lenient control measures for FMD than for rinderpest; notably, FMD-infected animals needed only to be isolated until their recovery, not slaughtered.

These measures failed to control an epidemic of FMD that, coincidentally, arose in 1869. Opinion was divided over whether FMD was a serious disease that merited stringent control measures, including a ban on live-animal imports, or whether it was largely inconsequential. The former view was held by upper-class farmers of pedigree stock and Tory members of parliament (MPs), whose own animals would increase in value should live animal imports be stopped. The

latter view was supported by Liberal and urban MPs, who were concerned about maintaining the meat supply to the working classes and the principle of free trade.

Opinion gradually settled on FMD being a serious disease. All livestock imports were banned in 1884 and, in 1892, an act of parliament ordered the slaughter of animals infected with FMD. Woods attributes this transformation to several factors, including the greater influence of pedigree breeders and Tory MPs; the greater susceptibility of new breeds of livestock developed in Britain at that time; and the growing importance of meat and milk to British agriculture, which meant that losses from FMD were felt more keenly. But the most important factor, Woods contends, was the legislation used to control the disease. She writes: "FMD-as-plague wasn't always 'out there' in nature, awaiting discovery by enlightened individuals. It was a new creation, a by-product of the processes involved in its control. And as this new vision of FMD grew in strength, its social origins were gradually obscured, until it came to be viewed as an incontrovertible fact of nature."

The control measures were successful and, after 1885, Britain experienced nearly 40 years with only a handful of FMD outbreaks. Woods thinks that the agency responsible for British farming acquired a taste for freedom

from FMD that, subsequently, it could not give up. Small outbreaks were easily controlled by the slaughter of infected animals and their contacts. Large outbreaks tested this approach to its limit, often with no sign of disease control in sight, despite the culling of large numbers of livestock. In all cases, however, wholesale slaughter remained the principal control method, and alternatives, such as letting the epidemic take its course or, since the 1950s, vaccination, were ignored. In Woods' view, this stubborn approach, which has cost the lives of millions of animals, is mistaken.

A Manufactured Plague is a delight to read. Too many popular-science books today seem to be little more than adverts for the erudition and breadth of knowledge of the author, or are the work of a non-specialist with a good idea. Woods, a graduate of both veterinary medicine and the history of medicine, has written a book that resounds with her depth of knowledge of the subject matter. This account of the history of FMD in Britain, and its political context, will be enjoyed by anyone interested in the disease, whether scientist or student, legislator or farmer.

Woods provides detailed accounts of most of Britain's FMD outbreaks over the past hundred years, but the one disappointment is the relatively short shrift given to the 2001 epidemic: it receives only six pages, compared with 20 or more for the epidemics of 1922–24 and 1967–68. It is the 2001 epidemic, more than any other, to which the concept of a 'manufactured plague' can be most readily applied. In 2001, fewer farms were affected than in the two aforementioned outbreaks, but 20–30 times as many animals were slaughtered. Woods alludes briefly to the role of mathematical models in directing this control policy but, regrettably, steers clear of detailed analysis. ■

Matthew Baylis is at the Institute for Animal Health, Pirbright Laboratory, Ash Road, Pirbright, Surrey GU24 0NF, UK.

Hitting the right note

Nature's Music: The Science of Birdsong

edited by Peter Marler & Hans Slabbekoorn
Elsevier: 2004. 504 pp. £49.95, \$74.95

Fernando Nottebohm

Nature's Music is a remarkable book in many ways — the breadth of its coverage, the blend of field and laboratory studies, and the balance between facts and speculation. It describes how the sounds produced by birds develop in individuals, what information they convey, how they are repre-



Sing when you're winning: the male western meadowlark uses song to defend its territory.

sented in the brain, how they fit in their natural settings, and how they have been studied. This may seem like a provincial subject, but because it touches on so many issues, the book reaches into much that is fundamental to the study of learning and animal languages.

Another factor that makes the book special is that it encompasses the life and work of two individuals. It was developed from presentations at a symposium to celebrate the life of Luis Baptista, recently deceased and a superb naturalist with a passion for birds and their vocalizations. He was one of the first to demonstrate the importance of social context in vocal learning in songbirds, and was the first to provide an experimental demonstration of vocal learning in hummingbirds.

The book is also about the discipline's father figure, Peter Marler, who, together with Hans Slabbekoorn, is the editor of *Nature's Music*. Marler has produced other good books, but the choreography of this one deserves comment. He wrote two of the 14 chapters, and the work and thoughts of Marler and his many disciples permeate much of the text. Each of the remaining 12 chapters was produced by a different author or group of authors. And each chapter includes short vignettes by a third tier of authors, who focus on single aspects of the larger story. This approach could have resulted in a cacophony of voices and styles, yet through first-class editing the book remains unified. The quality of figures is excellent, as are the two CDs that convert many of the

sounds illustrated as sound spectrographs back into an impeccably clean soundtrack. In all these ways, this book is a labour of love.

The book suggests that birdsong is not just about aesthetics, although it can charm female birds and human aficionados of both sexes. Birdsong, we learn, is a vehicle for cleverly encoded information about species and individual identity, about health and age, about the willingness to defend a territory, about genetic fitness, and about suitability as a mate. In addition, because birdsong is so easily recorded, quantified, altered and reproduced, it lends itself well to experiments on animal communication.

For historical reasons, the book does not give equal weight to the sounds of all birds, and there is a strong bias towards those that are learned. This aspect of vocal ontogeny offers parallels with vocal development in humans, and this is why birdsong has attracted so much basic research. Vocal learning in mammals is rare — it is found only in humans, some cetaceans and possibly some bats — but it occurs in half of all living birds, who acquire their song by imitating that of older conspecifics.

One of the most interesting chapters deals with the information conveyed by calls, rather than song. The calls that signal the presence of a predator are often similar for local resident species. Thus homeland security in nature benefits from a diversity of sentinels, each watching from its special vantage point and sounding the alarm in a universal code.

I was mesmerized in the 1960s when I heard Marler lecture on vocal communication in birds. This book brings back the charm of 'the good old days', recounted by Marler in the first chapter, and presents the excitement of the ensuing 40-year harvest. Who would have guessed that there was so much to come?

This book will make fine reading for all those drawn to birds and their songs, and will provide a sturdy backbone for courses on animal behaviour, animal communication and learning. Those who labour all day in concrete jungles or in the confines of a laboratory may find in this book an incentive to strap on the binoculars, step outside and follow nature's music. ■

Fernando Nottebohm is at the Field Research Center, Rockefeller University, 495 Tyrrel Road, Millbrook, New York 12545, USA.

More on birdsong

The Singing Life of Birds (Houghton Mifflin, \$28) is a personal account by Donald Kroodsma of the art and science of listening to birdsong. It comes with a CD of sample sounds.

Birdsong by Don Stapp (Scribner, \$24) is an accessible account of the history of birdsong research and the author's encounters with two leading birdsong researchers.

Science in culture

Womb with a view?

All is not as it seems in a television programme on the life of a fetus.

Martin Kemp

We tend to believe what we see, provided that the image obeys certain rules of internal consistency. Although we have plenty of evidence to reject the old adage that the camera never lies, particularly in this era of digital manipulation, we are perceptually prone to trust a representation that exudes an air of the 'real thing', especially if it has a photographic look. Illustrators therefore have an ethical imperative to use their skills in the service of a contract of trust with the viewer. With recent advances in computer graphics, these issues of trust have assumed a heightened urgency.

Film-makers have access to techniques that can render in compelling detail interactive hordes of warriors using just a few real actors. That's entertainment, they say. But exploiting computer-generated (CG) imagery in public science broadcasting is quite another matter.

Visual images can now convey the unseeable in brilliant colours and with wondrous spatial conviction, whether they are great arrays in the cosmos or molecular engineering. Such images are clearly a good thing in the communication of science. However, the viewer is frequently not told about the status of the images. When they relate to matters of considerable emotional and social importance, the stakes in the contract of trust can be huge.

On our screens we see a beautiful picture of a four-month fetus, yielding nothing in pink-appeal to a Raphael bambino. The fetus wheels slowly in space, apparently expressing feelings through its gestures and expressions. A voice-over intones some syrupy poetry, apparently composed by



the precocious infant through the adult medium of poet Roger McGough. We are told that the images come courtesy of new 'four-dimensional' (4D) scans (three dimensions plus real time).

This is what we were shown in the two-hour programme *Life before Birth*, made in Britain by Pioneer Productions and directed by Toby McDonald. The film was screened in Britain as *In the Womb* on the National Geographic channel on 11 March and on Channel 4 on 9 April.

There were some glimpses of relatively raw scans, but most of the spectacular visuals relied on animated models made by Middlesex-based company Artem. The fetuses were sculpted in wax, cast in silicon and hand painted. Animation specialists MillTV — better known for special-effects work in the film *Gladiator* — then set the models in motion. The skill and imagination behind the models were of the highest order, and the

results were seductive, visually and emotionally. We felt that we were eye witnesses to a beauty and conscious life previously unseen.

But at no stage was it clear what we were seeing. The credits named the companies responsible, but did not explain how the images were generated, and they were all implicitly accorded the same level of 'visual truth'.

Only on MillTV's website is the process made clear: "After months of research, courtesy of 4D ultrasound scans, medical books and pictures of mummified fetuses,

MillTV developed anatomically accurate CG recreations of month-four and month-seven fetuses." Each elaborate and laborious animation involved such methods as "multi-layering" for "shadowing, depth of field and colour correction flexibility".

In an area of medicine where public feelings run violently high, more honesty is required if the contract of trust with the viewer is to be honoured. I should like to propose a law and a consequent rule. The law is that the greater the skill available for making utterly convincing and seductive images, the greater is the power of potential deception. The rule is that the more sophisticated the techniques, the greater is the responsibility for openness in explaining how the images have been generated and where they stand in relation to the raw data.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and co-director of Wallace Kemp Artakt.

D. BARLOW/PIONEER PRODUCTIONS

An autistic look at animals

Animals in Translation: Using the Mysteries of Autism to Decode Animal Behaviour

by Temple Grandin & Catherine Johnson
Scribner/Bloomsbury: 2005. 368 pp.
\$25/£16.99

Marian Stamp Dawkins

There are two remarkable things about Temple Grandin. The first is that she has arguably done more than anyone else in the world to improve the welfare of animals in a practical way. Her major contribution has been to go into places that most of us would probably prefer not to think about — slaughterhouses — and imagine what it would be like to be an animal on its way to being killed. She has dramatically improved the welfare of these animals, not by making



An eye for detail: Temple Grandin believes her autism helps her to see things like a cow does.

any expensive modifications to the slaughter plants but by suggesting simple changes that cost nothing, such as removing a yellow coat hanging over a grey fence, or altering

the lighting to eliminate shiny reflections from a puddle. By removing stimuli that frighten cattle and cause them to stop and pile up on one another, the cattle move more easily, they don't slip and fall, and the use of electric goads is almost unnecessary. These things are all very simple and effective. It's just that no one had thought of them before.

The second remarkable thing about her is that she is autistic.

In *Animals in Translation*, Grandin argues that these two things are intimately connected. Her autism, she believes, gives her a remarkable insight into the way animals see the world. Animals, like autists, concentrate on detail. It is obvious to her that the yellow coat would be a scary stimulus to a cow, but the rest of us, concentrating on the bigger picture, would simply not realize unless it was pointed out to us. If Grandin's claim that her autism helps her to see the world through the eyes of other species sounds far-fetched, we have to remember her

R. WINARD

phenomenal success in making a difference in practice. Millions of animals are now moving through slaughterhouses more calmly because of what she has seen through her autistic eyes.

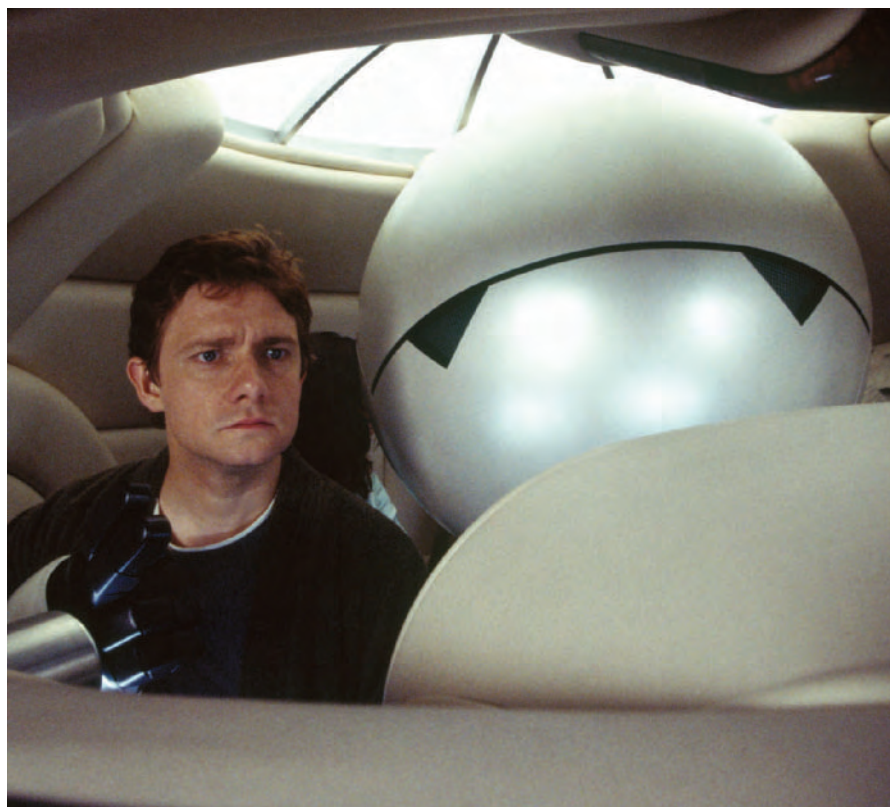
In some ways, this book is profoundly shocking, or at least it would be if it were not written by an autistic. "Autistic people are closer to animals than normal people are," she writes. "Autistic peoples' frontal lobes almost never work as well as normal people's do, so our brain function ends up being somewhere in between human and animal." Imagine anyone else saying that and getting away with it. But part of the power of this book is its innocence (her word too) and its genuine insights into a completely different way of thinking and seeing. Political correctness is not part of her world.

Grandin also states categorically that she doesn't have an unconscious and cannot repress unwelcome thoughts or emotions. This is one of many areas in which she has difficulty understanding normal people. She cannot understand why they deny what seems perfectly obvious to her, such as that something isn't going to work or shouldn't be said because it would offend people. Hearing her side of the story is a bit like being introduced to another culture by someone from that culture who has taken the trouble to find out which bits you are going to struggle to understand. Grandin straddles two worlds and, remarkably, can operate in both. If we want to understand her world, some of our preconceptions about what can and cannot be said will also have to be toppled. Being shocked out of our usual way of thinking is part of the process of understanding what it is like to be her.

At times, it is difficult to work out whether this is a book about animal behaviour with insights from autism, or a book about autism that uses animal behaviour to explain what it is like to be autistic. A major achievement of the book is that it is both, and it should be read as such. If you disagree with some of the things she says about animal behaviour or brain function, you keep going because of the fascinating insights you are gaining into her experiences as an autistic; and if you find yourself disagreeing with what she says about autism, you have the testimony of a unique human being about her life with animals.

Catherine Johnson has put Grandin's words into a form that is accessible to readers without any prior knowledge of either autism or animal behaviour. At the same time, people working in these areas will be given insights into the connections between these apparently diverse fields that will stick in the mind and change the way we look at both. ■

Marian Stamp Dawkins is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.



Guided tour: Arthur Dent and Marvin the Paranoid Android hitchhike through the galaxy.

Don't panic!

The Science of the Hitchhiker's Guide to the Galaxy

by Michael Hanlon
Macmillan Science: 2005. 256 pp.
£16.99, \$24.95

Joanne Baker

Your brief: to explain the science of Life, the Universe and Everything. It's quite a challenge. But Michael Hanlon pulls it off with wit, energy and style.

Timed to coincide with the release of the film adaptation of Douglas Adams' famous book series *The Hitchhiker's Guide to the Galaxy*, Hanlon's guide to the *Guide* takes the general reader on a grand tour of the outer reaches of modern scientific reality. Alien life, quantum physics and the history of the Universe are just as mind-boggling and weird now as they were to Adams in the 1970s, when he imagined Arthur Dent's escapades. Hanlon even manages to explain the unexplainable, such as the eye-popping shock of the total-perspective vortex.

Hanlon obviously enjoyed writing this book. It's not often that science writers get to rant about the non-existence of God, to explain the sudden appearance of a whale from a quantum fluctuation, or to ponder the genetic modification of animals to produce guilt-free meat. Adopting Adams' witty, punchy style, Hanlon's guide is a fun and vivid read. The science twinkles a little

more than usual in such a zany setting.

Although he tackles a wide range of cutting-edge topics with depth and authority, Hanlon has chosen the most obvious Hitchhiker destinations for his own scientific tour. The Restaurant at the End of the Universe prompts a discussion of the fate of the Universe; the babel fish yields a chapter on translation software; and time travel, parallel universes and black holes are well-trodden avenues. But when Hanlon does venture off-piste, he is a reassuring and insightful travelling companion, even if he often leaves the *Guide* behind. More references to it and amusing quotes could have added to the entertainment.

Readers familiar with the original *Hitchhiker's Guide* might have enjoyed more subtle tie-ins and a little more background about Adams himself, his peculiar ideas and influences. Hanlon briefly sets the context, but leaves such dialogue to others. The book also lacks comparisons with other contemporary science fiction. Aimed fair and square at the popular-science market, Hanlon's book may not satisfy die-hard science-fiction buffs. But the ghost of Adams is lurking in the pages.

See the film and buy the books. Don your striped jacket, bathrobe or spare head, and keep a towel handy. And above all — Don't Panic. With Hanlon's quirky book you are in safe hands. ■

Joanne Baker is in the Astrophysics Department, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX10 3RH, UK.

Natural symmetry

Directional inference: scientific convention applies conclusions from animal studies to humans but not the reverse, contradicting current evidence.

G. A. Bradshaw and
Barbara L. Finlay

Mickey Mouse, archie the cockroach, Snoopy, and Barney the purple dinosaur — just a few examples of creatures with minds and feelings like those of humans that richly populate the media. But in scientific circles, attributing human traits to animals is derisively dismissed as ‘anthropomorphism’. Biomedical science only permits, and critically depends on, inference from animals to humans, in the form of ‘animal models’. So in the direction of human–animal comparisons, popular culture seems to run contrary to science.

Are science and culture actually divergent here? No, not when you look more closely. Culture and science are based on the same simple model: an animal core painted with a veneer of human nature. Human ascendance, although only veneer-deep, is crucial to the model. Fanciful characters such as Barney and Snoopy, who sport a human psychic veneer on their respective dinosaur and canine effigies, are cultural heroes. In contrast, characters with the reverse attributes — Shakespeare’s unfortunate Bottom or Jeff Goldblum as *The Fly* — are punished or possessed, not improved.

Similarly, scientists have had no problem seeing the lessons of Harry Harlow’s experiments, in which monkeys were raised by artificial ‘surrogate’ mothers, for human orphanage practices. Neither have experiments using mice held dangling from their tails as models for human stress raised a scientific eyebrow, despite the tailless condition of *Homo sapiens*. But when it comes to applying conclusions about the human psyche to that of an animal, science draws a well-worn line.

The scientific reasoning behind this inferential line is now fading fast. Ideas about human–animal differences are being reshaped, and inference based on the veneer model of animal to human ascendance no longer fits current science. For years, most scientists with any biological training have discounted *scala naturae*, the idea of a progressive evolution from animals to the culmination of humans. We have long known that we share the same core but, more recently, our understanding of that core has been dramatically extended.

Today, continuity across species at a structural level is unquestioned in biomedical research, and the more we learn, the more



About face: new finds are forcing assumptions about our unique position in the animal kingdom to be revised.

apparent this continuity becomes. The same genes organize the body plan and fundamental structure-building and metabolic mechanisms across vertebrates and even invertebrates. Further, we have found that brain structures and organization develop in a highly coordinated fashion and that, across species, the principles of organization in the largest brains can be predicted from those in the smallest.

These structural similarities are echoed in behavioural patterns. Properties once thought uniquely human, such as culture, language, emotion, personality, are one by one being identified in species as varied as fish, sheep, rats, crows and even invertebrates. Not only can apes and birds design and use tools, but elephants can get post-traumatic stress, rodents can laugh, fish can suffer distress and, with a glance at its face, a sheep can assign another sheep to its correct position in the family tree and assess its emotional state.

Brain imaging has also played an important role in reshaping our views of the links between humans and animals, especially in relation to cognition, emotion and all ‘private’ mental states. This methodology has provided physical insight into mental states in humans, allowing direct comparison with mental states in animals. What was subjective has become objective, tractable and species-general. This implies that models of scientific inference are overdue for some serious rethinking to catch up and match scientific theory and data.

Science is not only composed of an ever-increasing number of facts; it also evolves through alterations in what are considered to be allowable methods and subjects of investigation, as the film *Kinsey* reminds

us. Pioneering scientist Alfred Kinsey argued in the 1940s and 1950s that although sexual behaviour can be dissected and catalogued, love cannot. Fifty years later, contrary to his predictions, science is doing just that: studying love, from the hormonal effects of ‘romantic disappointments’ in cichlid fish, and long-term pair bonding in voles, to the imaging of brain correlates of the ‘broken heart’ in humans. The mutual expansion of method and subject has thoroughly confused traditionally held views of species differences and the nature of emotions. Ironically, in our efforts to determine why we are so unique, we have discovered that we are not so different after all.

Certainly, true species differences remain, but we need a better model than a layer of paint. And even though understanding species commonality and diversity together requires some perceptual adjustment (particularly because so much of biology and psychology has been defined by the veneer model), alternative models are readily available given the convergence of theory and data from diverse fields.

Bringing symmetry to our inference-making will have important benefits that override any short-term impediments. Legitimizing a thoughtful form of anthropomorphism frees hypothesis generation from antiquated assumptions of human ascendance and, paradoxically, provides a much more coherent picture of both human and non-human species. Hopefully, for the next generation of schoolchildren, Lassie, Flipper and Skippy the bush kangaroo will seem as impossibly quaint as will progressive evolution and unidirectional inference to future scientists. Like Tyger in William Blake’s poem, symmetry may be fearful, but it is also fundamental. ■

G. A. Bradshaw is in the Environmental Sciences Graduate Programme, Oregon State University, Corvallis, Oregon, USA. Barbara L. Finlay is in the Departments of Psychology, and Neurobiology and Behavior, Cornell University, Ithaca, New York, USA.

FURTHER READING

Sapolsky, R. *A Primate’s Memoir: A Neuroscientist’s Unconventional Life Among the Baboons* (Scribner, 2001).

Finlay, B. L., Darlington, R. B. & Nicastro, N. *Behav. Brain Sci.* **24**, 263–308 (2001).

Gee, H. *Nature* **420**, 611 (2002).

Berridge, K. C. in *Handbook of Affective Sciences* (eds Davidson, R. J., Scherer, K. R. & Goldsmith, H. H.) Ch. 3 (Oxford Univ. Press, 2003).

Insulin trigger for diabetes

Matthias von Herrath

Type I diabetes occurs when the immune system destroys crucial cells in the pancreas. But what prompts the body to turn against itself so disastrously? It seems that insulin is the key.

Autoimmune disorders are caused by aberrant immune responses directed towards the body's own proteins (or 'autoantigens'). In type I diabetes, for example, immune cells called lymphocytes react against and destroy the β -cells in the pancreas of genetically susceptible individuals (Fig. 1). As these cells produce insulin — the hormone that helps to regulate glucose metabolism — their loss leads to the chaotic blood glucose levels that characterize diabetes and that have serious consequences. How the lymphocytes target β -cells has been unclear, but work from Nakayama *et al.*¹ and Kent *et al.*² in this issue (on pages 220 and 224) shows that insulin is itself an autoantigen that initiates the immune response leading to diabetes.

Although many autoantigens are thought to be involved in type I diabetes, conclusive evidence that any of these instigate or propagate autoimmune responses has remained elusive. Historically, however, insulin has been a good candidate. For example, lymphocytes isolated from 'non-obese diabetic' (NOD) mice — an animal model of type I diabetes — can recognize segments of the insulin molecule^{3,4}. In addition, autoantibodies against insulin are found in the blood of NOD mice and human diabetic patients, well before the clinical symptoms of the disease appear. The onset of clinical diabetes occurs after more than 90% of β -cells have been destroyed, but the insulin antibodies arise after only a small fraction of β -cells have been attacked, providing a useful early marker for the occurrence of autoimmunity.

Nakayama and colleagues¹ now provide convincing evidence that insulin is an essential autoimmune target in the initiation of diabetes. They stopped all natural insulin production in NOD mice by eliminating the genes that encode the two chains of insulin. So that the mice would not die, they reinstated a modified gene to produce insulin that is hormonally active but is not recognized by lymphocytes. Normally, NOD mice show an immune response against the β -cells — there are antibodies against the β -cells in their blood, and lymphocytes infiltrate into the pancreas, surrounding the β -cells before killing them. But the mice with the modified insulin showed no signs of any immune response against the β -cells, and they did not develop diabetes.

One question that remains is whether

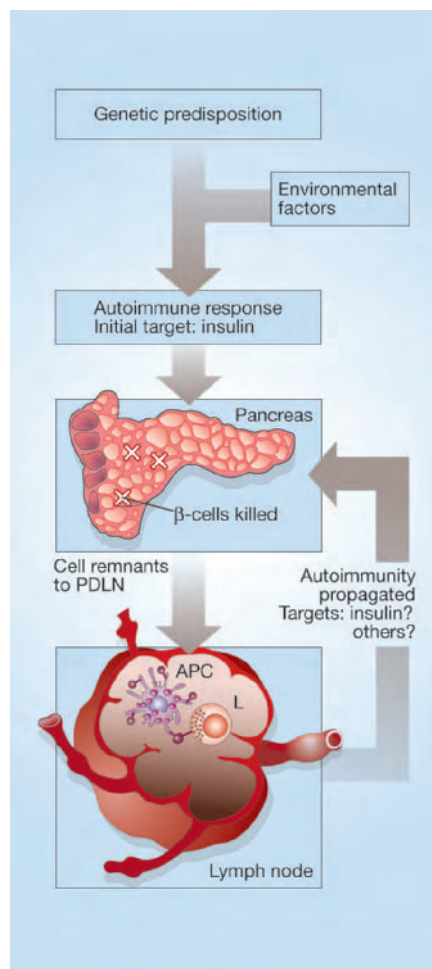


Figure 1 Development of type I diabetes. Genetic predisposition determines almost entirely whether a person will develop immune reactivity against insulin-producing β -cells in the pancreas. However, environmental factors such as nutrition and infections can have a major impact on whether type I diabetes manifests itself clinically; this occurs after 80–90% of the β -cells have been destroyed. The remnants of the β -cells are transported to the pancreatic draining lymph node (PDLN), where the ensuing autoimmune process is thought to be coordinated. Debris from the β -cells is picked up by antigen-presenting cells (APC) and 'displayed' to immune cells called lymphocytes (L), prompting the lymphocytes either to kill more β -cells or to signal further immune responses. Nakayama *et al.*¹ and Kent *et al.*² show that the initial immune trigger is insulin, although other immune targets could be involved later in the disease.

insulin is the crucial autoantigen only in the initiation of the disease, or whether other β -cell proteins become targets of the autoimmune response once the process has been set in motion. Lymphocytes do recognize several other targets as β -cell destruction progresses (by a process termed 'antigenic spreading')⁵. Notably, damped down expression of these antigens in β -cells, or elimination of the lymphocytes that recognize them, can lead to a significant reduction in the incidence of diabetes, but never to a complete absence of β -cell-specific autoimmune attacks^{6,7}. This suggests that although insulin is crucial for the initiation of β -cell destruction, other antigens might be involved in subsequent stages of the disease.

The relative importance of insulin over other candidate target antigens is underscored by Jaeckel and colleagues^{8,9}. They expressed insulin and glutamic acid decarboxylase (another protein proposed to be an autoantigen in type I diabetes) in the thymus of NOD mice. This meant that the immune system no longer reacted to these proteins. Tolerance to insulin led to a remarkable reduction in the occurrence of diabetes, but tolerance to glutamic acid decarboxylase did not.

How relevant are these findings for human diabetes? Kent and colleagues' results² support the concept that insulin is also an essential autoantigen in people. They isolated lymphocytes from the pancreatic draining lymph node of patients with type I diabetes, propagated them, and then analysed the proteins that the cells recognized. Remarkably, about 50% of the lymphocytes recognized a piece of the insulin A chain. By contrast, no healthy control subjects showed a similar accumulation of lymphocytes that recognize the insulin A fragment.

Lymphocytes come into contact with the insulin peptide as it is 'displayed' on the surface of so-called antigen-presenting cells; these cells collect protein fragments from dying β -cells before convening with the lymphocytes in the pancreatic draining lymph node. Kent *et al.* found that the cell-surface protein that binds to and displays the insulin A fragment on the antigen-presenting cells is encoded by a gene known to confer genetic susceptibility to diabetes.

Findings from Arif *et al.*¹⁰ also suggest that insulin could be the target of the

autoimmune response in humans, providing a potential candidate for treatment. Arif and colleagues found that insulin-reactive lymphocytes from diabetics appeared to have a destructive nature; they release molecules that are harmful to β -cells. Insulin-reactive lymphocytes from healthy individuals, however, seemed mostly to have a regulatory function, given the signalling molecules that they secrete. The idea that self-reactive lymphocytes can have different functions is therapeutically interesting in that it suggests two principal pathways to mitigate the effects of autoimmune lymphocytes: wipe out the aggressive cells, or ensure that more of the cells end up with a regulatory function^{11,12}. Indeed, recent work indicates that inducing such a change in the function of autoreactive lymphocytes from aggressive to regulatory might be possible, not only in mice, but also in some diabetic patients¹³.

So, we will not only have to learn which proteins and peptides are recognized by autoreactive lymphocytes, but we will also have to see whether this recognition elicits a regulatory or a destructive state. Dysregulation of such responses might ultimately lead to autoimmune diseases such as type I

diabetes, for example, if lymphocytes switch from a regulatory to an aggressive state. The identification of a high proportion of insulin-A-specific lymphocytes in human pancreatic lymph nodes gives us a valuable population of autoimmune lymphocytes to study, and highlights the importance of recovering cells directly from human target organs. ■

Matthias von Herrath is in the Immune Regulation Laboratory, La Jolla Institute for Allergy and Immunology, San Diego, California 92121, USA.
e-mail: matthias@liai.org

1. Nakayama, M. *et al.* *Nature* **435**, 220–223 (2005).
2. Kent, S. *et al.* *Nature* **435**, 224–228 (2005).
3. Wegmann, D. R., Norbury-Glaser, M. & Daniel, D. *Eur. J. Immunol.* **24**, 1853–1857 (1994).
4. Wong, F. S. *et al.* *Nature Med.* **5**, 1026–1031 (1999).
5. Tian, J. *et al.* *Immunol. Rev.* **164**, 119–127 (1998).
6. Lieberman, S. M. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 8384–8388 (2003).
7. Yoon, J. W. *et al.* *Science* **284**, 1183–1187 (1999).
8. Jaekel, E., Lipes, M. A. & von Boehmer, H. *Nature Immunol.* **5**, 1028–1035 (2004). Errata: **5**, 1190 (2004); **6**, 219 (2005).
9. Jaekel, E., Klein, L., Martin-Orozco, N. & von Boehmer, H. *J. Exp. Med.* **197**, 1635–1644 (2003).
10. Arif, S. *et al.* *J. Clin. Invest.* **113**, 451–463 (2004).
11. Homann, D. *et al.* *Immunity* **11**, 463–472 (1999).
12. von Herrath, M. G. & Harrison, L. C. *Nature Rev. Immunol.* **3**, 223–232 (2003).
13. The Diabetes Prevention Trial — Type I Study Group. *Diabetes Care* doi:10.2337/05.dc04-1693 (2005).

High-energy physics

An emptier emptiness?

Frank Wilczek

Temperatures similar to those reached an instant after the Big Bang can be created in collisions of gold atoms. The resulting fireballs may allow us a glimpse of a world that is more symmetrical than our own.

The concept that what we ordinarily perceive as empty space is in fact a complicated medium is a profound and pervasive theme in modern physics. This invisible, inescapable medium alters the behaviour of the matter that we do see. Just as Earth's gravitational field allows us to select a unique direction as up, and thereby locally reduces the symmetry of the underlying equations of physics, so cosmic fields in 'empty' space lower the symmetry of these fundamental equations everywhere. Or so theory has it. For although this concept of a symmetry-breaking aether has been extremely fruitful (and has been demonstrated indirectly in many ways), the ultimate demonstration of its validity — cleaning out the medium and restoring the pristine symmetry of the equations — has never been achieved: that is, perhaps, until now.

In a new paper, Cramer *et al.*¹ claim to have found evidence that — for very brief moments, and over a very small volume — experimentalists working at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory in New York have

vaporized one symmetry-breaking aether, and produced a more perfect emptiness. This pioneering attempt to decode the richly detailed (in other words, complicated and messy) data emerging from the RHIC experiments is intricate², and it remains to be seen whether the interpretation Cramer *et al.* propose evolves into a consensus. In any case, they've put a challenge on the agenda, and suggested some concrete ways to tackle it.

But what exactly is this underlying symmetry of nature that is broken by the aether? How is it broken, and how might it be restored? The symmetry in question is called chiral symmetry, and it involves the behaviour of quarks, the principal constituents of the protons and neutrons in atomic nuclei (among other things).

Chiral symmetry is easiest to describe if we adopt the slight idealization that the lightest quarks, the up quark (u) and down quark (d), are massless. (In reality their masses are small, on the scale of the energies in play, but not quite zero.) According to the equations of quantum chromodynamics (QCD), the theory that describes quarks and their interactions via the strong nuclear

force, the possible transformations among quarks are very restricted. One rule is that u-quarks and d-quarks retain their 'flavour' — that is, a u never converts into a d, nor a d into a u. Quarks also, like the more familiar photons, have an intrinsic spin. If the spin axis is aligned with the direction of motion, then the sense of the rotation defines a handedness, known as chirality, rather like a left- or right-handed screw. The two possible states of chirality of a quark, left and right, are essentially the same concept as left and right circular polarization for photons. The fundamental interaction between quarks and gluons, to which we ultimately trace the strong nuclear force, conserves chirality as well as flavour. Thus a u-quark with left-handed chirality (written u_L) never converts into a right-handed u_R , and so on. But these extra conservation laws, which follow from the symmetry of QCD's equations, are too good to be true. In reality, one finds that although the rule forbidding changes of flavour holds true, there is no additional conservation law for chirality — chiral symmetry is broken.

The accepted explanation for this mismatch blames a form of aether. The idea is that there is such a powerful attractive interaction between u_L -quarks and $\bar{u}_R$ -antiquarks (every quark has an antiquark with the opposite charge), and likewise between d_L -quarks and $\bar{d}_R$ -antiquarks, that the energy gained from their attraction outweighs the cost of creating the particles in the first place. Thus, perfectly empty space, devoid of quarks, is unstable. One can lower the energy of the vacuum by filling it with bound u_L - $\bar{u}_R$ and d_L - $\bar{d}_R$ pairs (and their antiparticles, $\bar{u}_L$ - u_R , $\bar{d}_L$ - d_R). Physicists call this process the formation of the chiral condensate. In the stable state that finally results, the conservation of chirality is rendered ineffective, as space itself has become a reservoir containing, for example, an indefinite number of u_L -quarks. (Because each u_L - $\bar{u}_R$ pair contains both a quark and an antiquark, net conservation of flavour still holds good.) This extraordinary picture has experimental consequences: the lightest strongly interacting particles, the π -mesons, can be identified as collective oscillations of the chiral condensate. This identification provides clues to the unusual properties and interactions of the π -mesons, notably their small masses compared with those of other strongly interacting particles.

At the RHIC, collisions between heavy ions — gold nuclei with a total of 197 protons and neutrons each — create a fireball in which temperatures exceeding 1.5×10^{12} kelvin are achieved (Fig. 1). Impressive evidence has accumulated that a qualitatively new state of matter has been created, a liquid-like plasma of quarks and gluons^{3–6}. Could something even more dramatic — a qualitative change in the properties of empty

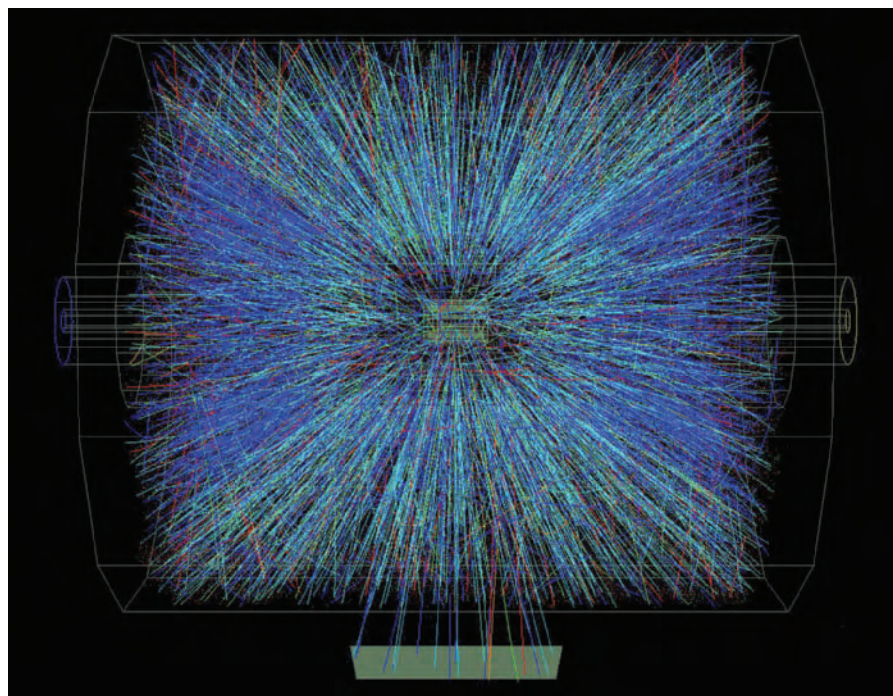


Figure 1 Gold dust. A side view of one of the first high-energy collisions captured by the Solenoidal Tracker of the STAR detector at the Relativistic Heavy Ion Collider (RHIC). The initial head-on collision of two gold ions, each consisting of a total of 197 protons and neutrons, occurs at the mid-point of the central tube (running across the image from right to left). The tracks indicate the paths taken by thousands of subatomic particles created in the fireball of energy set free in these collisions. Several layers of detectors, arranged concentrically around the central tube, and encased in a powerful magnet, allow the identification of these particles. (Courtesy of Brookhaven National Laboratory, STAR collaboration.)

space — be occurring as well? Theoretical calculations indicate that at such temperatures the pairs that make up the chiral condensate will break apart. When the condensate vaporizes, the full underlying chiral symmetry of QCD becomes operative. This change in the properties of ‘empty’ space last occurred throughout the Universe in the early moments after the Big Bang, when temperatures were as high as those reached in the RHIC fireball. This and similar vaporizations of other condensates at higher temperatures play an important role in modern cosmological thinking. Such an event might, for example, have triggered an epoch of inflation — a period of accelerated growth in which the horizon of the Universe expanded, temporarily, much faster than the speed of light.

Vaporization of the chiral condensate affords by far our best opportunity to access a phase transition of empty space in a controlled terrestrial experiment. The difficulty arises not so much in creating the necessary extreme conditions, but in reconstructing from the ashes available to us what happens during the initial stages of the experimentally created fireballs. Cramer *et al.*¹ use correlations between observed π -mesons to reconstruct properties of the medium through which they travelled. Previous models have had difficulty in dealing with these correlations, resulting in what has been

called the ‘HBT puzzle’⁷. Only by allowing for the possibility that the medium significantly alters the properties of the π -mesons, in the way expected if that medium were free of the chiral condensate^{8,9}, do Cramer *et al.* achieve a satisfactory fit to the data they consider. Thus, they may both resolve an old puzzle and open a new vista. Whether their model can be extended successfully to cover additional phenomena, and whether models based on other ideas can be equally successful, are questions sure to receive considerable attention in the near future.

Frank Wilczek is at the Center for Theoretical Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue 6-305, Cambridge, Massachusetts 02139-4307, USA.
e-mail: wilczek@mit.edu

1. Cramer, J., Miller, G., Wu, J. & Yoon, J.-H. preprint at www.arxiv.org/nucl-th/0411031 (2004).
2. Kolb, P. F. & Heinz, U. in *Quark Gluon Plasma* Vol. 3 (eds Hwa, R. C. & Wang, X.-N.) (World Scientific, Singapore, 2004); preprint at www.arxiv.org/nucl-th/0305084 (2003).
3. Adcox, K. *et al.* (The PHENIX collaboration) *Nucl. Phys. A* (submitted); preprint at <http://arxiv.org/nucl-ex/0410003> (2005).
4. Adams, J. *et al.* (The STAR collaboration) *Nucl. Phys. A* (submitted); preprint at <http://arxiv.org/nucl-ex/05010095> (2005).
5. Back, B. B. *et al.* (The PHOBOS collaboration) *Nucl. Phys. A* (in the press); doi:10.1016/j.nuclphysa.2005.03.084 (2005).
6. Arsene, I. *et al.* (The BRAHMS collaboration) *Nucl. Phys. A* (in the press); doi:10.1016/j.nuclphysa.2005.02.130 (2005).
7. Heinz, U. W. & Kolb, P. F. preprint at www.arxiv.org/hep-ph/0204061 (2002).
8. Son, D. T. & Stephanov, M. A. *Phys. Rev. D* **66**, 076011 (2002).
9. Son, D. T. & Stephanov, M. A. *Phys. Rev. Lett.* **88**, 202302 (2002).



100 YEARS AGO

A description of the large diamond found recently in the Premier Mine, Transvaal, is given in the *Geological Magazine* (April) by Dr. F. H. Hatch and Dr. G. S. Corstophine, with reproductions of four photographs which represent the diamond in its actual size from four different points of view... The stone is bounded by eight surfaces, four of which are faces of the original crystal, and four are cleavage surfaces, which are distinguished from the original octahedral faces by greater regularity and smoothness. For a large stone the crystal is of a remarkable purity... The stone, which has been named the Cullinan diamond, weighs 9600.5 grains troy, or 1.37 lb. avoirdupois; this is more than three times the weight of the largest diamond previously known.

ALSO:

In proposing the toast of “The Japan Society” at its annual dinner on May 3, Sir Frederick Treves referred to the medical and surgical ability of the Japanese. Nothing astounded him more, he said, in his recent visit to Japan than the way in which the Japanese have inquired into the medicine and surgery of the western world and the marvellous thing they are making of it... The Japanese have all the qualities of a surgeon. They have infinite patience and infinite tenderness. Sir F. Treves is confident that not many years hence there will be seen in Japan one of the most progressive schools of medicine the world has ever known. From *Nature* 11 May 1905.

50 YEARS AGO

In 1947, Evans and Guild described a technique for the quantitative extraction of earthworms. This consisted of treating a known area with a solution of potassium permanganate (1.5 gm. per litre) at the rate of 6.8 litres per sq. metre. The method was said to recover a high proportion of the total population... I found, however, that population estimates obtained by the permanganate technique were considerably lower than those suggested when a more laborious hand-sorting method was employed. In order to measure the relative efficacy of these techniques, permanent pasture on light soil of alluvial origin was sampled by both methods contiguously... While the fifty soil cores... produced 639.5 worms, the permanganate samples... produced only 350 worms. J. A. Svendsen
From *Nature* 14 May 1955.

Developmental biology

Asymmetrical threat averted

Eran Hornstein and Clifford J. Tabin

The somites are embryonic elements that give rise to the muscles, skeleton and some skin layers of the trunk. They form in a symmetrical fashion, but to do so they must be shielded from asymmetrical cues.

The human body looks deceptively symmetrical from the outside. In contrast to this external symmetry, our internal organs are in an asymmetrical, yet reproducible, arrangement. The heart is on the left and the lung next to it is smaller than that on the right. Within the abdominal cavity, the intestines form a clockwise loop; the stomach, pancreas and spleen are on the left, whereas the liver and gall bladder are on the right. But our body plan does not start out this way, and some body parts — the somites, for instance — must remain bilaterally symmetrical. A striking series of findings by Kawakami *et al.*¹ and Vermot and Pourquié², published on pages 165 and 215 of this issue, along with a complementary manuscript in *Science*³, teach us how somite formation proceeds in a bilaterally synchronized fashion (Fig. 1). Furthermore, Kawakami *et al.* and Tanaka *et al.*⁴, also writing in this issue (page 172), shed new light on how asymmetry, which is necessary for the physiological function of many organs, is first established in the embryo (Box 1, overleaf).

The left–right differences seen in many of our internal organs are rooted in a cascade of molecular asymmetries that is established during development (Box 1). Discoveries over the past decade have provided an increasingly detailed picture of how such left–right asymmetry is imposed on the developing embryo⁵. However, regulation of the alternative — bilaterally symmetrical morphogenesis — has not been given serious attention.

As the embryo begins symmetrically, it is

perhaps natural that researchers have intuitively viewed symmetry as a default state. For instance, this view has been applied to the somites, which give rise to tissues of the body wall, such as the musculature, skeleton and some skin layers. Somites are generated bilaterally in symmetrical pairs, with one somite of each pair on each side of the developing spinal cord. This process is driven by bilaterally symmetrical waves of gene expression⁶. The first pair of somites emerges at the anterior part of the embryo (next to the head), and more caudal pairs of somites form successively in an anterior–posterior sequence.

Kawakami *et al.*¹, Vermot and Pourquié² and Vermot *et al.*³ now show that somite formation can proceed in a bilaterally synchronized fashion only because somites are actively masked from information that would otherwise cause them to develop asymmetrically. Experiments in mammals, birds and fish all yield clues that somites are masked from left–right cues, and so are held in symmetrical register, by a signal that depends on retinoic acid (RA). When RA activity is experimentally attenuated, bilateral segmentation is taken out of phase.

This symmetry-generating mechanism is necessary because, to develop properly, somites need to be refractory to left–right signals that are present in their physical territory. In principle, somite cells could simply ‘ignore’ such cues by not expressing the receptive components needed to interpret them (if, for instance, they did not produce receptors for the relevant growth factors).

But what profoundly complicates things is that the very same left–right signalling cues are used to pattern somites along the anterior–posterior axis.

This problem of common signals being put to conflicting use finds its origin in a key principle that has emerged in modern molecular embryology. A surprisingly small number of molecular signals — such as Wnt, Notch, Hedgehogs, bone morphogenetic proteins and fibroblast growth factors (FGFs) — are repeatedly employed in different developmental settings, often in conjunction with distinct cofactors, to orchestrate differentiation and patterning. As in the composition of a musical piece, biological motifs are used over and over again, with changes of context and emphasis to generate new interpretations.

The anterior–posterior patterning of somites, for instance, is specified by the combined action of the Notch, Wnt, FGF8 and RA signalling pathways⁶. At the same time, the Notch, FGF8 and RA pathways are also instrumental in left–right determination⁵. This conflict makes it necessary to block the effect of side-specific signalling, to hold somitogenesis in synchronized and symmetrical register. That need is met, the new papers show^{1–3}, by an RA-dependent mechanism. More broadly, these new findings suggest that conflicting developmental pathways in general may be intermingled in a harmonious way through buffering mechanisms.

We have a lot to learn yet about such buffering mechanisms. In this particular case, it is still unclear exactly how RA mediates the masking of left–right signals. Might this molecule be differentially spread in the left and right somitic fields? On the basis of the expression patterns of messenger RNAs encoding RA-metabolizing enzymes, Kawakami *et al.*¹ suggest that the answer is no. However, a relevant enzyme might still be differentially regulated at a post-transcriptional level.

Another question concerns the relevant molecular targets of RA. RA is known to antagonize FGF8, which is generated at the posterior end of the embryo during somitogenesis. This antagonism is crucial for segmentation to proceed⁶, but it would be interesting to investigate whether the FGF8–RA interface also serves a role in buffering left–right signalling. An alternative protein that might buffer the system is Lefty-1, which is an important inhibitor of side-biasing information flow, although it acts at a stage downstream from the factors that impinge directly on the somites. Consistent with this possibility, the production of Lefty-1 is induced by RA and occurs in the midline of the embryo.

The new studies^{1–3} might also have clinical relevance. RA is produced from vitamin A in the body, so a maternal shortage of

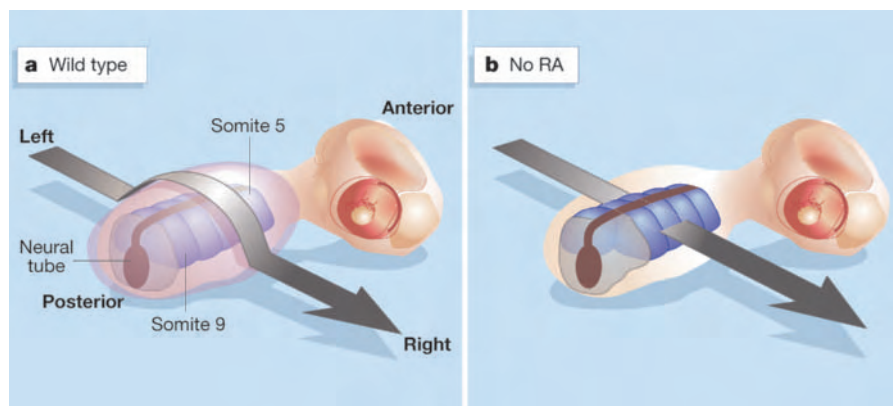


Figure 1 Maintaining symmetrical somitogenesis. A vertebrate embryo is shown from the back. a, The new papers^{1–3} show that, in wild-type embryos, a protective effect of retinoic acid (RA; pink field) masks the flow of left–right information (black arrow) to enable the symmetrical bilateral formation of new somites (blue). b, Blocking RA production exposes somites to left–right signals and takes their generation out of synchrony.

Box 1 Generating asymmetry

Over the past decade, much has been learnt about the genetic pathways that direct left–right asymmetry. A major remaining controversy has been the mechanism by which symmetry is first broken, in a reproducible orientation, in the early embryo.

In mammals, the initial symmetry-breaking event is believed to take place in the node — a small, transient pit at the anterior end of the primitive streak, near the middle of the early embryo. Thus, in mice, the first observable left–right difference is the directional rotation of microcilia (hair-like structures), which sets in motion a leftward flow of extracellular fluid across the node.

Two alternative mechanisms have been proposed for how fluid flow is sensed by the embryo: the unidirectional transport of a morphogen protein, and the

asymmetric deformation of mechanosensory cilia (the ‘two-cilia’ hypothesis; see ref. 4 and references therein). Tanaka *et al.*⁴ provide the first experimental support for the former model, showing that Sonic hedgehog (Shh) protein and retinoic acid are encapsulated, in a process that depends on fibroblast growth factor, into membrane-wrapped ‘nodal vesicular parcels’ (NVPs), which are asymmetrically transported by the flow across the node.

The biological significance of Shh transport remains unclear, however, as previous studies⁸ failed to detect any asymmetric readout of the Hedgehog pathway at the mouse node. Further, in the face of normal nodal hydrodynamics, mice with dysfunctional mechanosensory cilia exhibit defects in handedness⁵ that are consistent with the ‘two-cilia’ model, but are left unexplained in

the context of the NVP hypothesis. The discovery of the NVPs is intriguing, yet in its wake there are enough contradictions and open questions to stress the need for a unified interpretation that can explain all the existing data, and for an experimental setting that can definitively exclude one of those models.

Whichever way their action is interpreted, nodal cilia do seem to be at the root of mammalian symmetry breaking. In amphibians and birds, however, asymmetries are detected before the node forms, in some proteins of the extracellular matrix and in the activities of H^+/K^+ -ATPase channels and Notch proteins. This raises the possibility that the cilia-induced flow might be a mechanism unique to mammals.

However, Kawakami *et al.*¹, and Essner *et al.*⁹, show that ciliary movement and leftward flow are conserved early steps in

establishing left–right asymmetry, and verify that the activity of the microciliary motor protein (encoded by the *lrd* gene) is required for microcilia activity in zebrafish, as in mice. Yet Kawakami *et al.* also show that, in zebrafish, H^+/K^+ -ATPase and Notch act upstream of *lrd* and of the flow generated by microcilia. They conclude that the cilia may act to transduce the early left–right pattern that is initially established by H^+/K^+ -ATPase and Notch.

It remains to be determined how the H^+/K^+ -ATPase and Notch signalling systems produce a consistent left–right orientation, and how these earlier signals are connected to the latter effects of extracellular fluid flow at the node. Another question is whether these initiating steps are conserved in all vertebrates, or whether they have been lost in mammals, with the nodal flow being the true symmetry-breaking event. **E.H. & C.J.T.**

vitamin A during pregnancy might be associated with increased rates of fetal skeletal defects, such as hemivertebrae and scoliosis. The World Health Organization has found vitamin-A deficiency to be common in south Asia and Africa, suggesting that millions of pregnancies a year are carried by women with a vitamin-A shortage⁷. This astonishing number of pregnancies at potential risk might promote epidemiological studies of this theoretical correlation. ■

Eran Hornstein and Clifford J. Tabin are in the

Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

e-mail: tabin@genetics.med.harvard.edu

1. Kawakami, Y., Raya, A., Raya, R. M., Rodriguez-Esteban, C. & Izpisua Belmonte, J. C. *Nature* **435**, 165–171 (2005).
2. Vermot, J. & Pourquié, O. *Nature* **435**, 215–220 (2005).
3. Vermot, J. *et al.* *Science* doi:10.1126/science.1108363 (2005).
4. Tanaka, Y., Okada, Y. & Hirokawa, N. *Nature* **435**, 172–177 (2005).
5. Levin, M. *Mech. Dev.* **122**, 3–25 (2005).
6. Pourquié, O. *Int. J. Dev. Biol.* **47**, 597–603 (2003).
7. www.who.int
8. Zhang, X. M., Ramalho-Santos, M. & McMahon, A. P. *Cell* **105**, 781–792 (2001).
9. Essner, J. J. *et al.* *Development* **132**, 1247–1260 (2005).

Planetary science

Magnetic impact craters

David J. Dunlop

Aerial surveys of the Vredefort impact crater in South Africa suggest that it is only weakly magnetic. The rocks themselves tell a different story, but does this apply to giant impact basins on Mars?

Evidence about the history of Mars can be gleaned from magnetic surveys — hence the importance of data sent back by the Mars Global Surveyor in the 1990s¹. The magnetometer on the satellite measured much stronger magnetic fields over some parts of the southern highlands of the planet than fields at similar altitudes over Earth. But the fields are notably weak over the giant impact basins Hellas and Argyre (Fig. 1), suggesting that the dynamo in the martian core was inactive during the era of major

meteorite impacts 3.5–4 billion years ago^{2–4}. This conclusion has had a considerable influence on ideas about how Mars cooled, and when its mantle and inner and outer cores differentiated.

The hypothesis of demagnetized impact basins and all it implies about Mars’ evolution is called into question by Carporzen *et al.*⁵ on page 198 of this issue. They measure unusually strong magnetizations of bedrock samples in the giant Vredefort impact crater of South Africa, yet aerial measurements

of magnetic fields over the crater are lower than over surrounding areas. This paradox is explained by variations in the directions of sample magnetization vectors over distances of 10 cm or less. The strong, but spatially incoherent, magnetic signal of the bedrock is essentially randomized when viewed from larger distances. Meteorite craters can then seem to be magnetic or non-magnetic, depending on how close the magnetometer is to the source. Viewed from satellite altitudes of 100–400 km, martian impact basins would appear magnetically featureless if the magnetic vectors of their source rocks vary in direction over distances of a few kilometres or less.

The Vredefort crater is 2 billion years old and was originally 300 km in diameter. Vredefort is Earth’s oldest and largest impact structure, and Carporzen *et al.*⁵ argue that, as such, it is our best analogue for the giant martian impact basins. Moreover, other terrestrial impact structures, the Charlevoix and Slate Islands impact craters in Canada among them, also have strong but spatially dispersed sample magnetizations and low aeromagnetic signatures.

But what causes this unusual behaviour? Rocks typically become magnetized when they are cooled from a melt or other high-temperature conditions. This thermal remanent magnetization, like virtually all other mechanisms of magnetization, forms parallel to the magnetic field that acts during cooling of the magnetic minerals in the rock⁶. This parallelism is central to the success of

the palaeomagnetic method: a rock's remanent magnetism is a fossil compass that tracks the ancient magnetic field, recording movements of continents, sea floor and tectonic plates. To account for the magnetization of the Vredefort crater, some other magnetizing mechanism must have been at work during the meteorite impact.

Carporzen *et al.*⁵ suggest that highly charged plasmas produced in the impact created magnetic fields that lasted only a matter of minutes, but which were as much as 1,000 times stronger than Earth's magnetic field. The plasma currents that created the transitory fields would have varied on a scale of centimetres and are proposed to be the cause of the strong but randomly directed bedrock magnetizations. The magnetite that carries the bedrock magnetization is confined to planar deformation features in shocked quartz grains. The magnetic carrier certainly formed as a direct result of the shock: unshocked but otherwise similar rocks lack this magnetite. The orientation of the magnetite grains in the quartz cannot explain the random directions of magnetization — any external field in a constant direction would produce a distribution of magnetic vectors over a half-sphere, not over 360° as observed. The field itself must have had a random orientation, and remanent magnetization⁶ of the magnetite that crystallized during the brief existence of that field is responsible for the bedrock magnetization.

Supporting evidence comes from rocks that melted during the impact, which were also analysed by Carporzen *et al.*⁵ Unlike the shocked but unmelted bedrocks, these 'dyke' and 'vein' samples cooled for days before their magnetization became permanently fixed. By this time, the plasma fields had long vanished and the melt rocks acquired thermal remanent magnetizations as they cooled in Earth's field. Their magnetizations are much weaker than those of the surrounding bedrocks, but they are all in the same direction.

In drawing analogies between the Vredefort crater and giant impact basins on Mars, we must bear in mind the differences in scale, and in the types of rocks involved. Earth's continental crust, where most impact craters are preserved, tends to be granitic, rich in aluminium and silicon but poorer in magnesium and iron. The crust of the martian highlands is more basaltic⁷, similar to the composition of Hawaii, Iceland and the ocean floor. These rocks contain much more iron, and abundant iron-bearing magnetic minerals can form during cooling without the intervention of shock. The meteorite that formed the Vredefort crater was probably 10–20 km in diameter, but to produce a cavity 1,000 km wide, as Argyre initially was, requires an impacting body closer to 100 km in size. Shock-wave pressures are correspondingly larger and their effects are felt at greater depths.

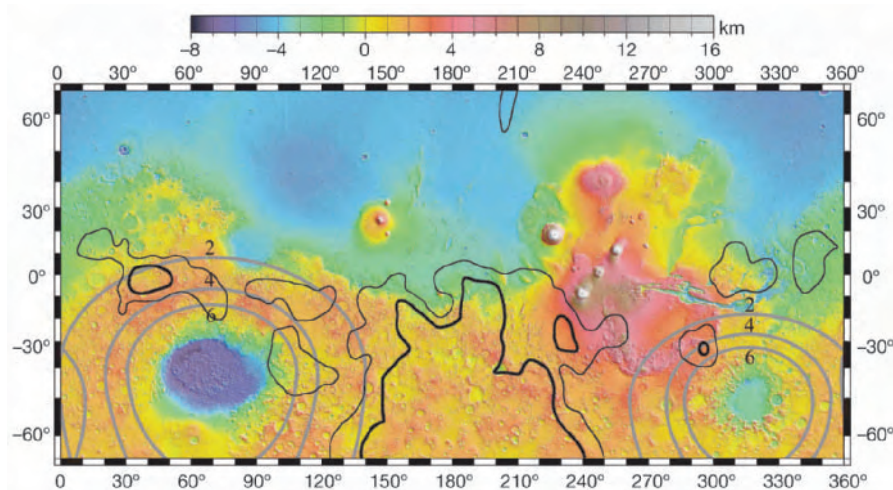


Figure 1 Magnetic lows over the Hellas (left) and Argyre impact basins on Mars. Superimposed on the topography are the peak shock pressures, in gigapascals, estimated to have been produced by the impact. The light and heavy magnetic-field contour lines are for 20 and 40 nanotesla, respectively, and were measured by the Mars Global Surveyor at about 400 km altitude. (Reproduced from ref. 3 with permission from the authors and the American Geophysical Union.)

The strongest magnetic fields measured by the Mars Global Surveyor require at least the upper 30 km of the crust to be magnetized^{8,9}. At these depths, shock and thermal blanketing would cause demagnetization of rocks that are already coherently magnetized — in addition to the new but randomly directed magnetization that plasma fields would create in crystallizing minerals. At the surface, the spatial coherence of bedrock magnetizations could be tested using an outcrop magnetometer on a future rover mission to Hellas or Argyre. It is noteworthy that some lunar rocks returned by the Apollo missions had strong and stable magnetizations whose directions varied by 60–120° between adjacent subsamples¹⁰ — these are a possible extraterrestrial example of the

suggested magnetization mechanism at Vredefort.

David J. Dunlop is in the Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada.

e-mail: dunlop@physics.utoronto.ca

1. Acuña, M. H. *et al.* *Science* **284**, 790–793 (1999).
2. Hood, L. L., Richmond, N. C., Pierazzo, E. & Rochette, P. *Geophys. Res. Lett.* **30**, 1281 (2003).
3. Rochette, P., Hood, L. L., Fillion, G., Ballou, R. & Ouladiaz, B. *Eos Trans. Am. Geophys. Un.* **84**, 561, 567 (2003).
4. Mohit, P. S. & Arkani-Hamed, J. *Icarus* **168**, 305–317 (2004).
5. Carporzen, L., Gilder, S. A. & Hart, R. J. *Nature* **435**, 199–201 (2005).
6. Dunlop, D. J. & Özdemir, O. *Rock Magnetism: Fundamentals and Frontiers* (Cambridge Univ. Press, 1997).
7. McSweeney, H. Y., Grove, T. L. & Wyatt, M. B. *J. Geophys. Res.* **108**, 5135 (2003).
8. Connerney, J. E. P. *et al.* *Science* **284**, 794–798 (1999).
9. Dunlop, D. J. & Arkani-Hamed, J. www.agu.org/pubs/pip
10. Fuller, M. *Rev. Geophys. Space Phys.* **12**, 23–70 (1974).

Sensory physiology

Brainless eyes

Rüdiger Wehner

The visual equipment of box jellyfish includes eight optically advanced eyes that operate with only a rudimentary nervous system. As they produce blurred images, their function remains an open question.

According to conventional wisdom, information-processing in visual systems is a hierarchical process¹. It starts at the level of the receptor layer, the retina, where raw sensory data are taken up from the outside world, and continues by transferring this information to increasingly higher centres in the brain. During this upstream process, exactly those features are extracted from the retinal response patterns that enable the animal to cope successfully with its ecological world.

The findings that Nilsson *et al.* describe

on page 201 of this issue² run counter to this view. A lowly animal inhabiting the tropical seas — a box jellyfish, or cubomedusa (Fig. 1) — is equipped with eight surprisingly sophisticated lens eyes of the camera-type, but there is no common brain behind them. In nearly every respect, these lens eyes resemble those of animals such as fish and cephalopods, but the 'central nervous system' behind the eyes consists only of a diffuse nerve net accompanied by a marginal nerve ring³.

Nilsson and colleagues' anatomical,

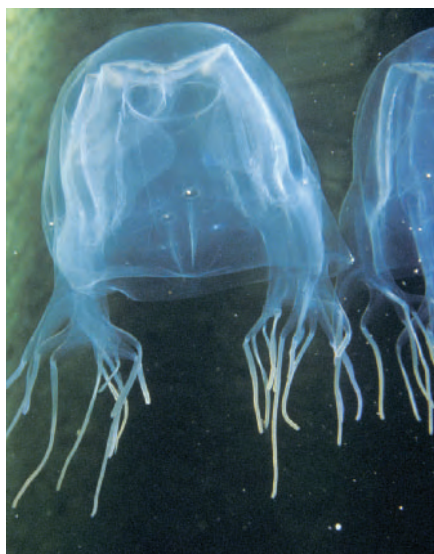


Figure 1 Eye see — but what? A box jellyfish of the genus *Chironex*. The species studied by Nilsson *et al.*² was *Tripedalia cystophora*, which can be bred from polyps in the lab.

optical (microinterferometrical) and modelling studies reveal two phenomena. The first is amazing, and the second peculiar. Amazingly, the tiny spherical lenses, which are only a tenth of a millimetre wide, are able to form sharp images, which are free of spherical aberration. This is due to a refractive index gradient within the lens — exactly as one would expect from optical theory, and exactly as is found in the much larger spherical lenses of fish and cephalopods⁴. Peculiarly, however, the focal plane of each jellyfish lens, and hence the sharp image formed by the lens, lies far behind the retina. This underfocusing leads to a blurred image, and thus to a loss of the fine visual detail that the lens is able to provide.

So why should evolution have produced highly sophisticated optics that have only poor resolution? Isn't higher resolution always better, irrespective of the visual function to be fulfilled? Obviously not. Cubomedusae are strong, agile swimmers and active predators³ living in near-shore habitats such as mangroves⁵. But how they are visually guided within these cluttered environments has remained elusive. Here we face the problem of what could be dubbed 'reversed neurobiology', analogous to the case of 'reversed genetics'⁶, in which the genes are known but their functions are not. We know what kind of visual cues the eyes of jellyfish are best at extracting, but not the visual tasks that the animals have to accomplish.

In insects, there are two examples in which the degradation of spatial resolution is a design feature of a particular visual subsystem known to serve a specific function. One is course stabilization, which involves horizon detectors that are built into small, single-lens eyes, the ocelli⁷. The other is skylight navigation, which is based on patterns of

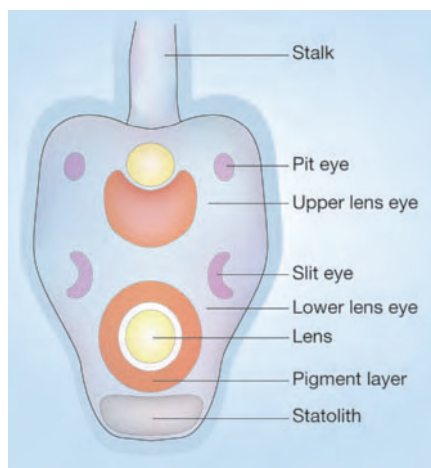


Figure 2 The rhopalium (sensory club) of a box jellyfish. In each of the four rhopalial located at the corners of the jellyfish's cube-shaped body there are two lens eyes² and four pigment-pit eyes (two upper pit eyes and two lower slit eyes). Each rhopalium hangs by a stalk from the rim of the medusa. The function of the crystalline concretion called the statolith might be to ensure that the rhopalium, and thus the eyes, are always oriented in the same way, irrespective of how the body of the medusa is tilted. (Diagram based on ref. 11.)

polarized light and mediated by a small, specialized part of the insect compound eyes⁸. But what are the jellyfish's eyes designed for?

This question is even more compelling, as Nilsson *et al.*² find that the photoreceptors of the jellyfish eyes possess wide and often complex (for example, asymmetrically shaped) receptive fields. In mammals, for instance, such complex receptive fields result from multi-level processing and hence are confined to higher cortical centres in the brain⁹. But in jellyfish they are generated by the optics of the lens and the position of the photoreceptors within the retina. This extreme case of peripheral filtering again hints at a very particular task that these eyes must

accomplish. It is further corroborated by the observation that, in each of the animal's four sensory clubs (or rhopalial), there are not only the two types of lens eye studied by Nilsson *et al.* — one looking upwards, the other horizontally — but also two other (simpler) types of eye (Fig. 2). Within these batteries of eyes, which exhibit quite some diversity of optical design, each type of eye could be specifically adapted to a particular aspect of the animal's lifestyle.

Nilsson and co-workers² provide us with one of the most dramatic examples that vision is not a general-purpose sense. Box jellyfish evidently have different eyes for different tasks, and have delegated such tasks, which are usually accomplished by neural processing, to the optics of the eyes. Furthermore, as the outputs of the eyes are channelled directly into the pacemakers for the swimming movements, and as these pacemakers¹⁰ are also located in the sensory clubs¹⁰, visuomotor processing occurs at an extremely peripheral level. Specialization of eyes for particular tasks and peripheral coding seem to go hand in hand — during the course of evolution, box jellyfish have clearly not had the need to feed the information provided by their total of 24 eyes into a central processing unit, or brain.

Rüdiger Wehner is in the Department of Zoology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.
e-mail: rwehner@zool.unizh.ch

1. Nicholls, J. G., Martin, A. R., Wallace, B. G. & Fuchs, P. A. *From Neuron to Brain* 4th edn (Sinauer, Sunderland, MA, 2001).
2. Nilsson, D.-E., Gislén, L., Coates, M. M., Skogh, C. & Garm, A. *Nature* **435**, 201–205 (2005).
3. Satterlie, R. A. *Can. J. Zool.* **80**, 1654–1669 (2002).
4. Land, M. F. & Nilsson, D.-E. *Animal Eyes* (Oxford Univ. Press, 2003).
5. Coates, M. M. *Integr. Comp. Biol.* **43**, 542–548 (2003).
6. Weissmann, C. *Trends Biochem. Sci.* **3**, 109–111 (1978).
7. Stange, G. *J. Comp. Physiol. A* **141**, 335–347 (1981).
8. Wehner, R. *J. Comp. Physiol. A* **189**, 579–588 (2003).
9. Schweigart, G. & Eysel, U. T. *Eur. J. Neurosci.* **15**, 1585–1596 (2002).
10. Satterlie, R. A. *J. Comp. Physiol.* **133**, 357–367 (1979).
11. Laska, G. & Hundgen, M. *Zool. Jb. Anat.* **108**, 107–123 (1982).

Granular media

Information propagation

Stefan Luding

The transmission of force through granular matter such as sand is a crucial consideration in certain applications. The behaviour observed depends on the particle interactions as well as on the length scale involved.

How is information propagated through granular matter? This is an essential question for researchers investigating the stability of buildings, silos and slopes — particularly for predicting failure and avalanches. Does propagation occur through specific ('easy') paths or along a wide front? Is the mechanism similar to that of sound- or light-wave

propagation? Is it elastic, or entirely different?

Goldenberg and Goldhirsch¹ provide answers to these questions on page 188 of this issue. Their numerical simulations show that, for short distances, forces in granular systems propagate much like waves, but that at longer distances an applied force causes an elastic-like deformation. Furthermore, the regime of wave-like behaviour becomes

smaller with increasing disorder and friction — thus, friction and disorder enhance elasticity.

In the general case of a point force applied to an elastic material in static equilibrium, the strength of the force decreases with distance. But even far away from the source, all material points are ‘aware’ of that force. This contrasts with wave propagation, for which practically no information is propagated outside characteristic ‘wave-fronts’ (rays), and also with diffusion, for which information is only ‘present’ behind the diffusive front.

The mathematical equations for elastic behaviour, waves and diffusion have different forms and thus different solutions. They must obey the basic laws of mass-, momentum- and energy-balance, like the Navier–Stokes equations of motion, which contain as special cases all of the above three types of behaviour. Thus, elastic solids and potential flows (for example, in water) behave elastically; sound is a propagating, wave-like mode; and heat is diffusive (Fig. 1).

What is the form of an equation that describes the propagation of information and, more specifically, the force, or stress, distribution in granular media? For granular materials such as sand, soil or snow, material disorder, anisotropy (direction dependence of properties) and friction have to be considered. In a fluid, the stress increases linearly with depth. But in a silo, the stress in the granular material saturates at a certain level that is independent of the depth, because static friction (not present in a fluid) causes the weight of the sand to be partially transferred to, and carried by, the silo’s walls². But still the question remains: how is the weight transported/propagated inside granular material?

When there are no side walls, as in the case of a conical sandpile, the situation is even less clear. One would expect the stress to be maximal under the apex (the highest point). Surprisingly, some experiments³ reveal that the stress actually shows a local minimum — a ‘dip’ — right under the apex. This observation is one of the reasons behind the revival of interest in the problem of force (or stress) propagation in granular materials. Although the dip cannot be explained using the simplest elasticity theory⁴, wave-like models readily predict it^{5,6}. But the dip can also be reproduced by including anisotropic elasticity, for instance⁴. Invoking special boundary conditions, such as a deflection of the supporting ground⁴ or a rough floor⁷, allows elastic models also to mimic the dip. Furthermore, it turns out that the existence of the dip depends on the structure of the sandpile and the way it was constructed^{8–10}. Thus, the problem also concerns the interplay between extrinsic boundary effects and intrinsic granular-material behaviour.

Recent laboratory experiments^{8–13} have probed the more fundamental response

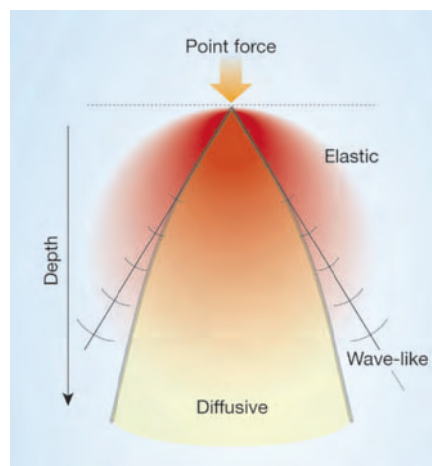


Figure 1 Three modes of information propagation in two dimensions. In response to a point force, an elastic mode (red) transmits the information (force) throughout the entire medium, with the force decaying, as indicated by the colour intensity. In contrast, diffusively propagated information remains behind a rather sharp ‘diffusion front’ (lines enclosing the yellow area). Wave-like propagation occurs along the straight lines (black) fanning out from the source.

of a system to a point-like source of perturbation. But they have not led to a common conclusion. Some find wave-like behaviour whereas others do not, or observe mixed behaviour^{8,11}; in some very small systems, even diffusive-like behaviour seems to exist¹². These discrepancies can be due to the effects of the finite size of the laboratory systems¹³, and the idealized particles (discs or spheres) or the two-dimensional nature of some of the experiments. Another possibility is that all types of behaviour (elastic, wave-like and diffusive) are present and that other details, as mentioned above, determine which one is observed¹⁴. Although most experimental observations can be explained with numerical and theoretical models^{1,4–6,15–21}, they all depend on certain basic assumptions and again do not provide conclusive answers.

Goldenberg and Goldhirsch¹ investigate a numerical model of two-dimensional slabs composed of disc-shaped grains subjected to a local, downward, vertical force. This approach has a notable advantage over physical experiments — one can systematically vary the material properties, control the system parameters and gather the desired information. In principle, the approach allows the individual effects of different particle sizes and shapes, packing structure, anisotropy and friction to be studied. The authors observe elastic behaviour in granular media by including the effects of sufficiently strong disorder and friction in their simulations^{1,20,21}. In contrast, wave-like propagation is observed in assemblies of frictionless particles¹⁷ and/or in regular, lattice-like structures¹⁵, or only on rather

small scales of 10–20 particle diameters^{1,14}. Thus, even though wave-like behaviour is observed for short distances in laboratory experiments, classical elastic-like behaviour is obtained for the larger length scales of structures such as dams, building foundations and silos. Furthermore, Goldenberg and Goldhirsch find that both disorder and friction — relevant for most realistic, large-scale systems — reduce the short-range wave-like regime and enhance elasticity. But as their conclusions are based on two-dimensional numerical simulations of discs, they will need to be checked in three dimensions, on larger systems and with less-idealized, non-spherical particles.

The puzzling question about the form of information propagation in quasi-static granular media seems to be solved, in that elastic-type behaviour occurs. Nevertheless, it may be that the local or large-scale ‘fragile’ (nearly unstable) states of granular matter (connected to wave-like information propagation) are also relevant. On rare occasions, perhaps such factors are responsible for silo collapse or foundation failure.

Many other properties of granular matter are still not understood. Examples are dynamic sound propagation, jamming (exceedingly slow dynamics), ratcheting (a system’s response under cyclic loading, such as trucks on roads or trains on rails), creep and phenomena such as quicksand. Granular matter offers plenty of other puzzles and challenges.

Stefan Luding is in the Particle Technology Group in Chemical Technology, Applied Sciences Faculty, Technical University of Delft, Julianalaan 136, 2628 BL Delft, the Netherlands.

e-mail: s.luding@tnw.tudelft.nl

- Goldenberg, C. & Goldhirsch, I. *Nature* **435**, 188–191 (2005).
- Janssen, H. A. Z. *Ver. Dt. Ingen.* **39**, 1045–1049 (1895).
- Smid, J. & Novosad, J. *Proc. 1981 Powtech Conf., Inst. Chem. Eng. Symp.* **63**, D3V 1–12 (1981).
- Savage, S. B. in *Physics of Dry Granular Media* (eds Herrmann, H. J., Hovi, J.-P. & Luding, S.) 25–95 (Kluwer, Dordrecht, 1998).
- Wittmer, J. P., Claudin, P., Cates, M. E. & Bouchaud, J.-P. *Nature* **382**, 336–338 (1996).
- Bouchaud, J.-P., Cates, M. E. & Claudin, P. *J. Phys. I* **5**, 639–656 (1995).
- Luding, S. *Phys. Rev. E* **55**, 4720–4729 (1997).
- Geng, J. *et al. Phys. Rev. Lett.* **87**, 035506 (2001).
- Spannuth, M. J., Mueggenburg, N. W., Jaeger, H. M. & Nagel, S. R. *Granular Matter* **6**, 215–219 (2004).
- Reydellet, G. & Clément, E. *Phys. Rev. Lett.* **86**, 3308–3311 (2001).
- Geng, J., Reydellet, G., Clément, E. & Behringer, R. P. *Physica D* **182**, 274–303 (2003).
- Da Silva, M. & Rajchenbach, J. *Nature* **406**, 708–710 (2000).
- Stone, M. B. *et al. Nature* **427**, 503–504 (2004).
- Howell, D., Behringer, R. P. & Veje, C. *Phys. Rev. Lett.* **82**, 5241–5244 (1999).
- Otto, M., Bouchaud, J.-P., Claudin, P. & Socolar, J. E. S. *Phys. Rev. E* **67**, 031302 (2003).
- Blumenfeld, R. J. *Phys. A* **36**, 2399–2411 (2003).
- Tkachenko, A. V. & Witten, T. A. *Phys. Rev. E* **60**, 687–696 (1999).
- Moukarzel, C. F., Pacheco-Martinez, H., Ruiz-Suarez, J. C. & Viales, A. M. *Granular Matter* **6**, 61–66 (2004).
- Liffman, K., Nguyen, M., Metcalfe, G. & Cleary, P. *Granular Matter* **3**, 165–176 (2001).
- Goldenberg, C. & Goldhirsch, I. *Phys. Rev. Lett.* **89**, 084302 (2002).
- Goldhirsch, I. & Goldenberg, C. *Eur. Phys. J. E* **9**, 245–251 (2002).

Obituary

Stanley J. Korsmeyer (1950–2005)

When Stan Korsmeyer was in the eighth grade and looking forward to playing high-school sports, he was told by the school coach that he would have to grow six inches before he would be allowed on any of the teams. Being determined but nonetheless pragmatic, Korsmeyer decided to focus his energies in an alternative direction — raising hogs. Raise hogs he did, and at age 14 Korsmeyer became the youngest person in the history of the Illinois State Fair to show the Grand Champion pair of Hampshire hogs. Two old-time Illinois farmers were heard discussing the competition. One asked, “Who showed the Grand Champions this year?” The other replied, “Just some kid.” It was not “just some kid”. It was Stan Korsmeyer, who even then dedicated himself to and excelled at whatever he did.

Korsmeyer was born on 8 June 1950, on a farm among the cornfields of southern Illinois, and he planned to be a veterinarian. However, as an undergraduate at the University of Illinois in Urbana, he became interested in curing people instead of livestock. In 1976 he earned an MD at the University of Illinois in Chicago. A move to the University of California, San Francisco, allowed him to continue his medical training and to meet his wife-to-be Susan, an oncology nurse. This move also introduced Korsmeyer to sailing and fishing, two passions that continued throughout his life.

Korsmeyer's next move was to the opposite coast (in part to continue his proximity to an ocean), to the National Cancer Institute in Bethesda, Maryland, where, as a fellow in molecular oncology, he trained with leading cancer researchers Thomas Waldmann and Philip Leder. Korsmeyer continued his career as an independent researcher at the NCI, and then as an investigator of the Howard Hughes Medical Institute, first at Washington University in St Louis and later at the Dana-Farber Cancer Institute and Harvard Medical School in Boston.

Korsmeyer opened new doors to the understanding and treatment of cancer. He co-discovered the oncogene *Bcl-2*, which when overexpressed can lead to follicular lymphoma. His studies provided evidence that *Bcl-2* was the first member of a new category of cancer-causing genes — genes that work not by driving cells to proliferate, but rather by preventing them from dying via a process known as programmed cell death, or apoptosis. By causing cells that normally die to



Trailblazer in the understanding and treatment of cancer

survive instead, elevated expression of *Bcl-2* can allow those cells to accumulate additional mutations and become cancerous.

Korsmeyer also discovered other members of the *Bcl-2* gene family and found that they too regulate cell death. Whereas members such as *Bcl-2* protect cells from dying, others, including *Bax*, *Bad* and *Bid*, cause cells to die. Korsmeyer proposed a rheostat model for the regulation of cell death, in which the crucial ‘life versus death’ decision made by each cell is determined by the balance between the activities of antagonistic survival and killer genes. He also discovered that *Bcl-2* family members are associated with mitochondria, the power plants of cells, and thus implicated mitochondria in the process of apoptosis.

Korsmeyer's studies helped to establish that many human disorders — including lymphomas and other cancers — can be caused by a misregulation of apoptosis. In short, he made crucial contributions to the elucidation of the molecular-genetic pathway that controls cell survival and cell death and to the discovery that the

misregulation of cell death has a major role in human disease.

That a deficit in apoptosis can cause tumours implied that cancers might be treated by reactivating apoptosis and causing tumour cells to self-destruct. Korsmeyer's research elegantly demonstrated that cancer cells can be induced to die either by blocking anti-cell-death genes such as *Bcl-2*, or by activating pro-cell-death genes such as *Bax*. As a consequence, the *Bcl-2* family has become a promising target for anticancer drugs. Korsmeyer himself worked directly to develop new treatments for cancer based on his knowledge of apoptosis. Although we met frequently at scientific meetings, it was in the context of our serving as consultants to a start-up biotechnology company focused on apoptosis that I got to know Korsmeyer best. His science and advice helped to make *Bcl-2* a major target, and as a consequence a drug that blocks *Bcl-2* action is currently under development at a major pharmaceutical company.

Korsmeyer's drive to develop an anticancer drug and to cure cancer patients was tempered by his concern about the risks of clinical trials. “I would never want to be involved in the trial of a drug that I wouldn't be willing to test on my own sons,” he once told me. Those sons, Evan and Jason, as well as his wife Susan, were at the centre of his life, and his statement about drug trials reflected Korsmeyer's deep empathy for others.

Korsmeyer had a farm boy's work ethic, dedication, energy and modesty. He also had an unshakeable determination, as well as a great sense of humour, a joyous and immediately recognizable laugh, enormous personal warmth, an optimism about people, and an infectious and constantly upbeat approach to life. I never knew anyone who did not like him. He cared enormously about the young scientists who trained with him; he was a superb mentor and seeded the biomedical world with his scientific offspring. He received many well-deserved prizes and honours for his discoveries.

Stan Korsmeyer, a non-smoker, died on 31 March of lung cancer. His death is a tragic loss to science, medicine and humanity. He may never have grown the six inches demanded by his high-school coach, but he was nonetheless a giant — as a scientist, a mentor, a friend and a human being.

H. Robert Horvitz

H. Robert Horvitz is an investigator of the Howard Hughes Medical Institute at the Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.
e-mail: horvitz@mit.edu

Genetics

Mutation and speech impairment

Am. J. Hum. Genet. **76**, 1074–1080 (2005)

Studies of a particular family with developmental speech difficulties have implicated a mutation in a gene known as *FOXP2* in such disorders. In a wider study, Kay D. MacDermot and colleagues have now identified another mutation elsewhere in the same gene that also seems to impair speech.

The previously identified defect in *FOXP2* is a 'missense' mutation that alters the amino-acid sequence of the resulting protein, a gene-transcription factor. Members of the family who carry this mutation have problems controlling the precise muscle movements required for speech, a condition known as developmental verbal dyspraxia.

MacDermot *et al.* screened the entire *FOXP2* gene in 49 volunteers with verbal dyspraxia. In one person, a four-year-old boy, they found a 'nonsense' mutation that causes the protein product to be truncated and to lack several crucial domains, including a DNA-binding site. The boy had language skills far below those expected for his age. Moreover, his mother and sister carried the same mutation, and also reported language problems, whereas his father had an apparently normal gene and normal speech.

Evidently, there may be a range of defects in *FOXP2* that result in incorrect speech development.

Michael Hopkin

Aeronautics

That leaking feeling

Appl. Phys. Lett. **86**, 174105 (2005)

The dangers posed to manned space flight by small air-leaks in the skin of a spacecraft — caused by space debris or a meteorite impact — are obvious. Stephen D. Holland *et al.* present a method that is suitable for detecting and localizing millimetre-size holes in metal structures under space conditions.

Conventional leak-detection strategies used on Earth — such as recording of characteristic ultrasound whistles with directional microphones — do not work so well in space, where the air carrying the noise escapes into the vacuum. Holland and colleagues developed a technique that relies instead on mounting a few vibration sensors directly on the inner surface of a pressure vessel. A computer algorithm allows the authors to 'feel' the leak by correlating the individual sensor signals. These experimental correlations are compared with dynamically computed correlations, so that the leakage

can be spotted reliably and reproducibly.

Several sensor configurations were tested in the laboratory. Although the algorithm occasionally reported spurious leaks, it did not miss a single real one. The authors believe that their technique is now ready for terrestrial testing on space hardware.

Andreas Trabesinger

Cell biology

Calcium connection

J. Cell Biol. **169**, 435–445 (2005)

Calcium ions regulate much of what happens in a cell, and a search has long been on to identify the mechanism underlying the phenomenon known as store-operated calcium entry. Jack Roos and colleagues have come up with what appears to be a crucial protein in the pathway.

In several cell types, when intracellular calcium stores are depleted, an ion channel opens up on the cell surface that allows calcium to enter. Working with insect cells, Roos *et al.* used the technique of RNA interference to pinpoint an essential player in this process — a protein called Stim. 'Knockdown' of the mammalian version of the protein, STIM1, reduced store-operated calcium entry in three different lines of human cells. Interestingly, overexpression of STIM1 produced only a modest increase in calcium influx, so it looks as if the protein is not itself the channel.

STIM1 is a single-pass transmembrane protein with a calcium-binding component, and it is present both at the cell surface and

in the membranes of intracellular organelles. These characteristics make it a good bet in connecting store depletion to calcium entry.

Lesley Anson

Medical imaging

Vessel in sight

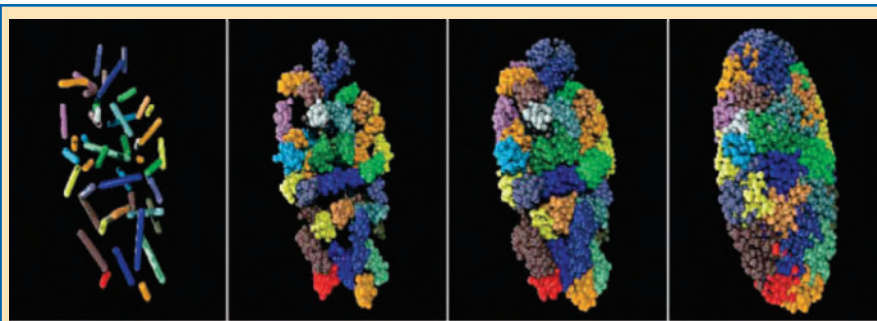
Chem. Commun. doi:10.1039/b502347e (2005)

How can clinicians check for the growth of new blood vessels around a tumour? The process of 'angiogenesis' cannot easily be visualized using the conventional technique of magnetic resonance imaging (MRI) — hence Anouk Dirksen and colleagues' interest in developing a better 'contrast agent'.

To enhance the quality of an MRI scan, scientists use contrast agents containing the paramagnetic metal gadolinium wrapped in an organic molecule that targets a specific tissue type — cancer cells, for example. Dirksen *et al.* have produced a contrast agent that ties an atom of gadolinium to a cyclic peptide, cNGR, which is known to bind specifically to CD13 — an enzyme that is overexpressed in growing blood vessels.

By adding the protein avidin to the contrast-agent mix, the authors also find that up to four molecules of cNGR can bind to a single molecule of CD13. Accumulating gadolinium atoms around a blood vessel in this way should further enhance the clarity of an MRI image, and the authors now plan to test their supramolecular contrast agent in mice.

Mark Peplow



Molecular biology

A new dimension to DNA

PLoS Biol. **3**, e157 (2005)

An innovative technique for three-dimensional imaging offers geneticists a fuller picture of what happens inside the cell nucleus. Andreas Bolzer *et al.* have developed a fluorescent staining method that allows them to visualize all chromosomes at key points in the cell cycle, while preserving the shape and position of each 'chromosome territory'.

Working with human fibroblast cells cultured from a skin biopsy, the team investigated the positions of all chromosomes in the nuclei of cells that had stopped dividing. They found that small chromosomes in non-dividing fibroblasts stuck

close to the centre of the nucleus, whereas larger ones cosied up to the outer edge.

Different cell types show different chromosome arrangements, and although nuclei vary in shape — flat ellipsoidal in fibroblasts versus spherical in lymphocytes, for instance — geometry doesn't seem to explain the variations, as computer models using geometrical rules to fill the nuclear space (pictured) put small chromosomes at the outside. Gene density might be the answer, though, as gene-poor regions tended to be close to the nuclear envelope, whereas gene-rich segments were farther inside. The authors predict that examining three-dimensional images will enable researchers to link the position of genes within the nucleus to their subsequent expression.

Roxanne Khamis

Self-reproducing machines

A set of modular robot cubes accomplish a feat fundamental to biological systems.

Self-reproduction is central to biological life for long-term sustainability and evolutionary adaptation. Although these traits would also be desirable in many engineered systems, the principles of self-reproduction have not been exploited in machine design¹. Here we create simple machines that act as autonomous modular robots and are capable of physical self-reproduction using a set of cubes.

A physical system is self-reproducing if it can construct a detached, functional copy of itself — by definition, this will also be capable of self-reproduction. Self-reproduction differs from self-assembly², in which the resulting system is not able to make, catalyse or in some other way induce more copies of itself. These phenomena have been of interest since the early days of computation^{3,4}, but have been examined mostly in abstract^{5,6} and simulated^{7–9} systems.

The self-reproducing machines demonstrated here are essentially modular robots¹⁰. Their modules have electromagnets that selectively weaken and strengthen connections, determining where the structure breaks and joins. Each module is a 10-cm cube, split into two halves along the (111) plane (Fig. 1a). One half of the cube can swivel relative to the other half in increments of 120°, each time cycling three faces of the cube. Connected cubes can both form and change into arbitrary arrangements (Fig. 1b). The cubes are powered through the baseplate and transfer data and power through their faces. The control of the machine is distributed among the modules: a microcontroller in each module executes a motion schedule governed by time and contact events. (For details, see supplementary information.)

In order to self-reproduce, a machine requires a supply of material. We supplied the modular robots with cubes that were manually replenished at two 'feeding' locations. The four-module robot (Fig. 1c; for movie, see supplementary information) was able to construct a replica in 2.5 min by lifting and assembling cubes from the feeding locations. Because the replica is as large as the original, the replica reconfigures itself to assist in its own construction. A three-module robot is able to self-reproduce in just over 1 minute (see supplementary information). Other reproducing forms are also possible with these modules⁸.

Self-reproduction of a physical machine has previously been achieved in two dimensions using tumbling wooden tiles¹¹ and a machine comprising four different components that are assembled by following tracks¹². In neither case, however, was it clear how to

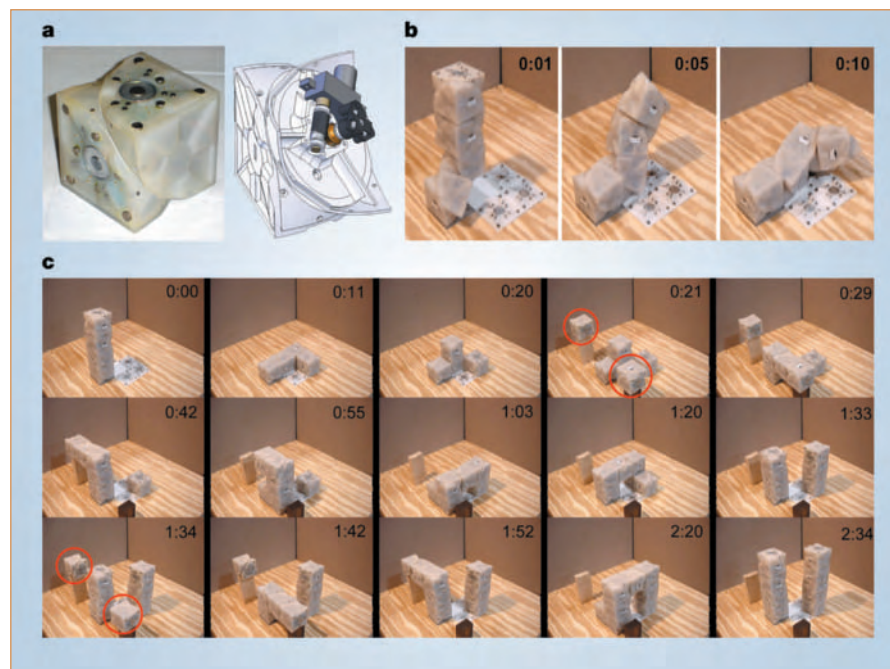


Figure 1 Self-reproduction of a four-module robot. **a**, Basic module, with an illustration of its internal actuation mechanism. **b**, Snapshots from the first 10 s showing how a four-module robot transforms when its modules swivel simultaneously. **c**, Sequence of frames showing the self-reproduction process, which spans about 2.5 min and runs continuously without human intervention, apart from the replenishing of building blocks at the two 'feeding' locations (circled in red). (For movie, see supplementary information.)

scale the process to more complex systems, short of redesigning the 'atomic' components.

In our demonstration, we use a modular substrate in which arbitrarily complex self-reproducing machines can be constructed. We circumvent the long-standing hurdle of what counts as self-replication by suggesting that self-replicability is not a binary property that a system either possesses or not, but is a continuum dependent on the amount of information being copied. This factor can be measured by comparing the log probability of a machine spontaneously appearing in an environment to the log probability of it appearing, given that one instance already exists. This factor can be computed precisely for some well-defined formal systems¹³ and approximated for others. For example, an abstraction of Penrose's replicating tiles¹¹ yields a factor between zero (not self-replicating) and log 2.

Even without calculating absolute numbers, systems can be ranked by comparing properties that affect this factor, such as the number of basic building blocks used compared with the number of building-block types and their complexity. This factor is exceedingly high for animals, which have about 10^{20} amino-acid combinations of roughly 20 amino-acid types, but is very low for our robots (four modules of one complex type). This view allows us to quantify,

compare and systematically improve the processes of self-reproduction. It is possible, for example, that self-reproducing machines composed of many identical microscale modules would improve this factor.

Although the machines we have created are still simple compared with biological systems, they demonstrate that mechanical self-reproduction is possible and not unique to biology. This design concept could be useful for long-term, self-sustaining robotic systems in emerging areas such as space exploration and operation in hazardous environments, where conventional approaches to maintenance are impractical.

Victor Zykov*, **Efstathios Mytilinaios†**, **Bryant Adams‡**, **Hod Lipson***

Departments of *Mechanical & Aerospace Engineering, †Computer Science, ‡Mathematics and §Computing & Information Science, Cornell University, Ithaca, New York 14853, USA
e-mail: hod.lipson@cornell.edu

1. Sipper, M. & Reggia, J. A. *Sci. Am.* **285**, 26–35 (2001).
2. Jackman, R. J., Brittain, S. T., Adams, A., Prentiss, M. G. & Whitesides, G. M. *Science* **280**, 2089–2091 (1998).
3. Von Neumann, J. in *Essays on Cellular Automata* (ed. Burke, A. W.) 4–65 (Univ. of Illinois, Illinois, 1970).
4. Freitas, R. A. & Merkle, R. C. *Kinematic Self-Replicating Machines* (Landes Bioscience, Georgetown, Texas, 2004).
5. Langton, C. G. *Physica* **10**, 134–144 (1984).
6. Lohn, J. D. & Reggia, J. A. *IEEE Trans. Evol. Comp.* **1**, 165–178 (1997).
7. Butler, Z., Murata, S. & Rus, D. *Distrib. Auton. Robot. Syst.* **5**, 37–48 (2002).

8. Efstathiou, M., Marcus, D., Desnoyer, M. & Lipson, H. *Designed and Evolved Blueprints For Physical Self-Replicating Machines in Proc. Ninth Int. Conf. Artificial Life 15–20* (Boston, Massachusetts, 2004).
 9. Rubenstein, M., Krivokon, M. & Shen, W.-M. in *Proc. IROS 2004* 2661–2666 (Sendai, Japan, 2004).
 10. Yim, M., Zhang, Y. & Duff, D. *IEEE Spectrum* **39**, 30–34 (2002).
 11. Penrose, L. S. *Sci. Am.* **200**, 105–114 (1959).
 12. Chirikjian, G. S., Zhou, Y. & Suthakorn, J. *IEEE/ASME Trans. Mechatron.* **7**, 462–472 (2002).
 13. Adams, B. & Lipson, H. in *Lecture Notes in Computer Science* Vol. 2801 (eds Banzhaf, W. et al.) 1–9 (Springer, Germany, 2003).
- Supplementary information accompanies this communication on Nature's website.
- Competing financial interests: declared none.

Botany

A record-breaking pollen catapult

The release of stored elastic energy often drives rapid movements in animal systems^{1,2}, and plant components employing this mechanism should be able to move with similar speed. Here we describe how the flower stamens of the bunchberry dogwood (*Cornus canadensis*) rely on this principle to catapult pollen into the air as the flower opens explosively^{3–5}. Our high-speed video observations show that the flower opens in less than 0.5 ms — to our knowledge, the fastest movement so far recorded in a plant.

Cornus canadensis grows in dense carpets in the vast spruce-fir forests of the North American taiga. As bunchberry flowers burst open, their petals rapidly separate and flip back to release the stamens (Fig. 1). During the first 0.3 ms, the stamens accelerate at up to $24,000 \pm 6,000 \text{ m s}^{-2}$ (2,400g), reaching the high speed ($3.1 \pm 0.5 \text{ m s}^{-1}$) necessary to propel pollen, which is light and rapidly decelerated by air resistance (terminal velocity, $0.12 \pm 0.03 \text{ m s}^{-1}$ (mean $\pm$ s.e.m.); $n=7$). The pollen granules are launched to an impressive height of 2.5 cm (range, 2.2–2.7 cm; $n=5$), which is more than ten times the height of the flower: from this height, they can be carried away by the wind. (For methods and movies, see supplementary information.)

Petals open independently of stamen activity, moving out of their way within the first 0.2 ms (Fig. 1). Petals attain a maximum speed of $6.7 \pm 0.5 \text{ m s}^{-1}$, accelerating at up to $22,000 \pm 6,000 \text{ m s}^{-2}$ (or 2,200g). The process of petal opening and pollen launch

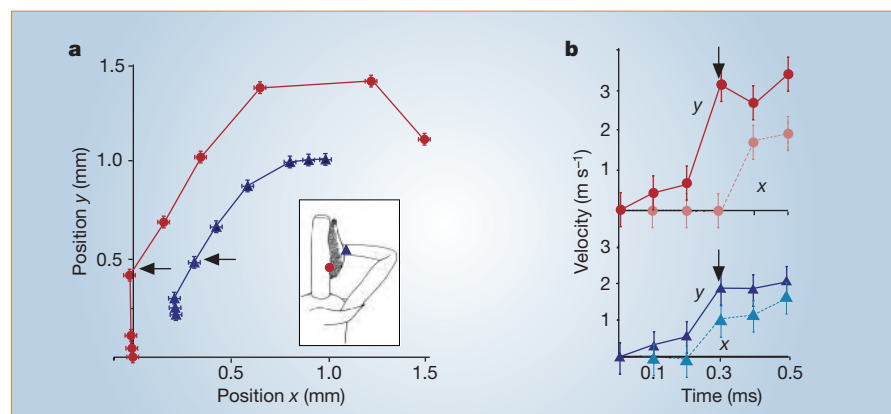


Figure 2 Dynamics of floral explosion. **a**, Coordinates x and y of positions of the filament tip (blue triangles) and anther tip (red circles), plotted at 0.1-ms intervals. Inset, a single stamen; points used to plot positions are indicated. Arrows, stamen positions just before pollen release. **b**, Coordinates x and y of velocity components of the anther (top) and filament (bottom) as a function of time, derived from the first six points in **a**. Arrows, velocity just before pollen release. Error bars represent uncertainty in measurements from **a**, propagated as random errors.

in bunchberry plants occurs faster than the opening of *Impatiens pallida* fruits (2.8–5.8 ms, $n=3$, see supplementary information); the snap of venus flytraps (*Dionaea muscipula*; 100 ms)⁶; the leap of froghoppers (*Philaenus spumarius*; 0.5–1.0 ms)¹; or the strike of the mantis shrimp (*Odontodactylus scyllarus*; 2.7 ms)².

As in these other organisms^{1,2,6}, rapid movements in bunchberry flowers rely on stored mechanical energy. Physiological processes, which take about a millisecond for each enzymatic reaction⁷, are not required for the explosion itself. We find that the flowers will open even when the stamen filaments have been crippled by treatment with sodium azide. But the flowers do not open if their turgor is reduced: dehydration of flowers with sucrose decreases the extent of opening, although subsequent rehydration allows them to open fully (results not shown). Turgor pressure is therefore required in the production of mechanical energy for explosive flower opening.

Bunchberry stamens are designed like miniature medieval trebuchets — specialized catapults that maximize throwing distance by having the payload (pollen in the anther) attached to the throwing arm (filament) by a hinge or flexible strap (thin vascular strand connecting the anther to the filament tip). This floral trebuchet enables stamens to propel pollen upwards faster than would a simple catapult. After the petals open, the bent filaments unfold, releasing elastic energy. The tip

of the filament follows an arc, but the rotation of the anther about the filament tip allows it to accelerate pollen upwards to its maximum vertical speed, and the pollen is released only as it starts to accelerate horizontally (Fig. 2).

The rapid opening of the self-incompatible⁸ bunchberry may enhance cross-pollination in two ways. First, when insects trigger flower opening, the pollen released sticks to their body hairs until it is transferred to an adhesive stigma. The force required to open flowers (0.1–0.5 mN) favours large pollinators (bumbees, for example) that move rapidly between inflorescences; it effectively excludes smaller, less mobile visitors such as ants. Second, pollen from flowers that open by themselves may be carried by wind currents. Indoors, pollen is transported over 22 cm (more than 100 times the size of the flower) and outdoors, in the presence of a steady wind, pollen can move farther than a metre. Exploding flowers enhance insect pollination and may allow wind pollination, adding to growing evidence that flowers often use multiple pollination mechanisms^{9,10}.

Joan Edwards*, Dwight Whitaker†, Sarah Klionsky*, Marta J. Laskowski‡

Departments of *Biology and †Physics, Williams College, Williamstown, Massachusetts 01267, USA
e-mail: joan.edwards@williams.edu

‡Biology Department, Oberlin College, Oberlin, Ohio 44074, USA

1. Burrows, M. *Nature* **424**, 509 (2003).
2. Patek, S. N., Korff, W. L. & Caldwell, R. L. *Nature* **428**, 819–820 (2004).
3. Lovell, J. H. *Bull. Torrey Bot. Club* **25**, 382–390 (1898).
4. Marie-Victorin, F. *Flore Laurentienne* (Imprimerie de la Salle, Montreal, 1935).
5. Mosquin, T. *Can. Field-Nat.* **99**, 1–6 (1985).
6. Forterre, Y., Skotheim, J. M., Dumais, J. & Mahadevan, L. *Nature* **433**, 421–425 (2005).
7. Voet, D. & Voet, J. G. *Biochemistry* 2nd edn (Wiley, New York, 1995).
8. Barrett, S. C. H. & Helenurm, K. *Can. J. Bot.* **65**, 2036–2056 (1987).
9. Kelly, D., Ladley, J. J., Robertson, A. W., Edwards, J. & Smith, D. C. *Nature* **384**, 615 (1996).
10. Kearns, C. A., Inouye, D. W. & Waser, D. N. *Annu. Rev. Ecol. Syst.* **29**, 83–112 (1998).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

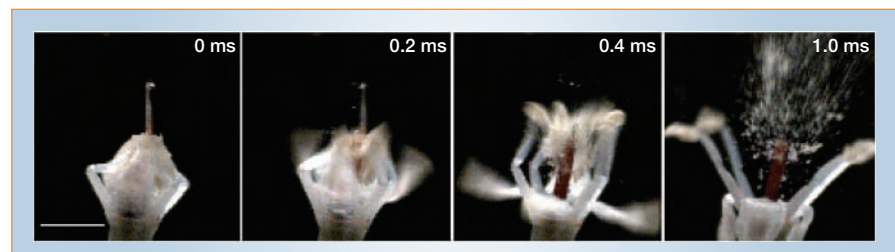


Figure 1 Bunchberry flower opening, recorded on video at 10,000 frames per second. Time elapsed is indicated. First frame shows a closed flower with four petals fused at the tip, restraining the stamens. Blur represents the distance moved in 0.1 ms. Scale bar, 1 mm.

Retinoic acid signalling links left–right asymmetric patterning and bilaterally symmetric somitogenesis in the zebrafish embryo

Yasuhiko Kawakami*, Ángel Raya*, R. Marina Raya, Concepción Rodríguez-Esteban & Juan Carlos Izpisua Belmonte

Gene Expression Laboratory, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

* These authors contributed equally to this work

During embryogenesis, cells are spatially patterned as a result of highly coordinated and stereotyped morphogenetic events. In the vertebrate embryo, information on laterality is conveyed to the node, and subsequently to the lateral plate mesoderm, by a complex cascade of epigenetic and genetic events, eventually leading to a left–right asymmetric body plan. At the same time, the paraxial mesoderm is patterned along the anterior–posterior axis in metamerical units, or somites, in a bilaterally symmetric fashion. Here we characterize a cascade of laterality information in the zebrafish embryo and show that blocking the early steps of this cascade (before it reaches the lateral plate mesoderm) results in random left–right asymmetric somitogenesis. We also uncover a mechanism mediated by retinoic acid signalling that is crucial in buffering the influence of the flow of laterality information on the left–right progression of somite formation, and thus in ensuring bilaterally symmetric somitogenesis.

The body plan of vertebrates has distinct left–right (LR) asymmetries in the disposition of internal organs. Cells and tissues are instructed as to their left or right identity at very early stages of embryo development. A complex cascade of epigenetic and genetic mechanisms ultimately results in the directional transfer of laterality information to the regions adjacent to the embryo node by the late gastrulation and early somitogenesis stages. The identity of such mechanisms is beginning to be understood and involves signalling by the Wnt, fibroblast growth factor (FGF) and Notch pathways, among others (reviewed in refs 1–5). The LR asymmetric placement of internal organs contrasts with (and is concealed by) the bilateral symmetry of the musculoskeletal and dermal outer layers of the body wall. This bilaterally symmetric arrangement arises from the synchronized segmentation of the paraxial mesoderm into aligned pairs of somites along the anterior–posterior (AP) axis, a process that also takes place in the vicinity of the embryo node and is controlled by the Wnt, FGF and Notch pathways, among others (reviewed in refs 6–9). Thus, during a sizeable developmental window, the same signalling pathways convey both symmetric and asymmetric information for bilaterally synchronized somitogenesis and LR asymmetric patterning, respectively. This seeming paradox poses a challenge that is currently unexplained.

Here we show that the establishment of the LR axis in zebrafish embryos is controlled by two parallel mechanisms, one that depends on H^+/K^+ -ATPase activity and acts very early in development (before the start of zygotic transcription), and a second mechanism akin to the nodal flow of mouse embryos, which takes place in Kupffer's vesicle (KV) during early somitogenesis and depends on the expression of *left–right dynein* (*lrd*) and functional cilia. We also identify a mechanism mediated by retinoic acid (RA) signalling that provides a molecular link between the development of AP and LR axes and is essential to buffer the influence of LR asymmetric information on the synchronized and hence bilaterally symmetric elongation of the AP axis. Our findings illustrate the existence of higher levels of coordination between the establishment of embryonic axes and organ patterning¹⁰.

Early steps of LR patterning in zebrafish

To address a possible role of H^+/K^+ -ATPase activity in zebrafish LR patterning, we incubated embryos with the specific inhibitor omeprazole from the 1–2-cell stage until bud stage (zebrafish embryo staging is described in ref. 11). LR patterning defects were observed, as demonstrated by reversed heart looping (31% ($n = 98$) versus 1% ($n = 90$) in control embryos; Fig. 1a, g). Moreover, expression of the *nodal*-related gene *southpaw* (*spaw*) and its target *pitx2*, normally restricted to the left lateral plate mesoderm (LPM^{12–14}; Fig. 1b, c), seemed randomized in embryos treated with omeprazole (44% ($n = 45$) and 43% ($n = 82$), respectively; Fig. 1h, i). Similar results were obtained with SCH 28080, a reversible inhibitor of H^+/K^+ -ATPase activity (Table 1 and data not shown). The fact that these treatments induced high percentages of LR alterations when performed before midblastula transition¹⁵ (Table 1) indicates that the H^+/K^+ -ATPase relevant for LR patterning in the zebrafish is maternally transcribed, as it is in *Xenopus* embryos¹⁶. H^+/K^+ -ATPase α immunoreactivity was readily detected in 2-cell-stage embryos, and was subsequently localized to all blastomeres with no apparent asymmetries up to the 1k-cell

Table 1 LR patterning defects induced by manipulations during development

Developmental stages	Abnormal phenotypes (%; n)			
	DMSO	SCH 28080	DAPT	<i>lrd</i> -MO
1–32 cells	0; 78	27; 84	2.8; 144	23; 86
32–1k cells	1.0; 102	23; 123	4.0; 75	–
High to 50% epibody	0; 88	9.6; 83	2.9; 136	20; 120
50–75% epibody	0.9; 113	4.0; 75	3.9; 129	–
75% epibody to 1 somite	0; 98	1.1; 94	20; 179	–
1–10 somites	0; 95	0.6; 158	0; 108	–

Zebrafish embryos were incubated in the presence of dimethylsulphoxide (DMSO; 1%), SCH 28080 (100 μ M) or DAPT (100 μ M) during the developmental windows indicated, followed by extensive washing with embryo medium; for *lrd*-MO experiments, embryos were injected at the 1-cell stage or at the 1k-cell stage. Embryos were allowed to develop and heart looping was scored 48 h post-fertilization. Values are percentages of abnormal heart looping phenotypes (leftward or absent) and the total numbers of embryos tested.

stage (Supplementary Fig. S1). Thus, in the zebrafish, as in the chick¹⁶, the role of H^+/K^+ -ATPase activity during LR patterning seems to depend on post-translational differences, rather than those at the transcriptional level as in *Xenopus*¹⁶. Our findings uncover a very early role of H^+/K^+ -ATPase activity in the control of LR patterning in zebrafish embryos, which, to our knowledge, represents the earliest known step in the cascade of laterality information in this species.

In chick embryos, it has been shown that the difference in H^+/K^+ -ATPase activity results in increased levels of extracellular Ca^{2+} on the left side of the node, which in turn results in a local increase in Notch activity necessary for correct LR patterning¹⁷. Downregulation or upregulation of Notch signalling in mouse^{18,19} and zebrafish¹⁹ embryos, respectively, also results in LR patterning defects, indicating a possible evolutionarily conserved role of Notch activity in the LR information cascade. To address whether Notch activity is necessary for LR patterning in the zebrafish, we incubated embryos with the γ -secretase inhibitor DAPT (*N*-[*N*-(3,5-difluorophenacetyl-L-alanyl)]-(*S*)-phenylglycine *t*-butyl ester)^{20,21}. Incubation with DAPT from bud to early somitogenesis stages resulted in alterations in LR patterning at the level of visceral LR asymmetries (24% (*n* = 46); Fig. 1m), and in the expression of *spaw* and *pitx2* (38% (*n* = 50) and 32% (*n* = 41), respectively; Fig. 1n, o). The fact that these treatments perturbed LR patterning most frequently when performed after 75% epiboly and before the 1-somite stage (Table 1) is in concordance with the proposed role of Notch signalling in the establishment of the LR axis in mouse^{18,19} and chick¹⁷ embryos, and shows that in zebrafish embryos Notch signalling acts downstream of H^+/K^+ -ATPase in the cascade of laterality information.

A role for cilia in zebrafish LR patterning

In the mouse embryo, the initial LR symmetry-breaking event is thought to take place in the node and to be mediated by the directional leftward flow of extracellular fluid generated by the rotation of monocilia in nodal cells²². It has also been proposed that a similar mechanism operates in vertebrate embryos other than the mouse, on the basis of the existence of cilia and the expression of *Lrd* in the node of *Xenopus*, chick and zebrafish embryos²³, although experimental confirmation of this suggestion is lacking. In contrast, LR asymmetries exist in *Xenopus* and chick embryos before node formation (reviewed in ref. 24), and our results described above with inhibition of H^+/K^+ -ATPase and Notch activities indicate that this is most probably true of the zebrafish as well. Thus, if *Lrd* or other motor proteins have a function in LR patterning in these embryos, it is likely to be the amplification of upstream LR information rather than its generation *de novo*. We addressed this possibility by downregulating *lrd* translation using morpholino-modified antisense oligonucleotides (MO). Injection of *lrd*-MO into 1-cell-stage embryos resulted in a high frequency of LR patterning defects, as evaluated by reversed or failed heart looping (23% (*n* = 86); Fig. 1s) and altered expression of *spaw* and *pitx2* in the LPM (40% (*n* = 67) and 40% (*n* = 36), respectively; Fig. 1t, u). As expected, this treatment did not alter *lrd* expression in KV (Fig. 1d, v).

We next explored the possibility that, as in mouse embryos²⁵, the role of *lrd* in zebrafish LR patterning was related to cilia function and the generation of a directional fluid flow. For this purpose, we observed the cilia of cells lining KV (the zebrafish equivalent to the mouse node²³) in control embryos and *lrd* morphants. Whereas cilia were readily visible and extended along the entire bottom of KV in

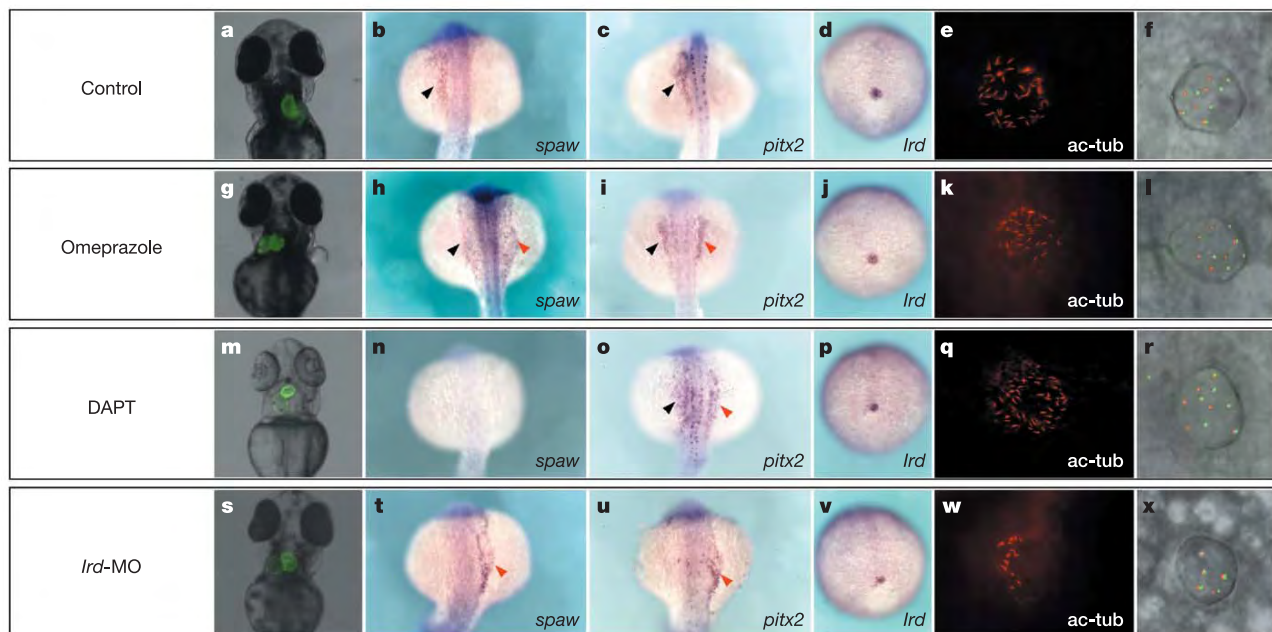


Figure 1 Cascade of LR asymmetric information in the zebrafish. **a–f**, Control zebrafish embryos have a rightward looping of the heart, revealed by green fluorescence in the *mlc2a*-GFP transgenic line (**a**), left-sided expression of *spaw* (**b**) and *pitx2* (**c**) in the lateral plate mesoderm (black arrowheads), expression of *lrd* in KV (**d**), numerous cilia in KV, as evaluated by acetylated tubulin immunoreactivity (ac-tub; **e**), and a net leftward flow inside KV, as evaluated by the movement of injected fluorescent beads (**f**). **g–x**, Inhibition of H^+/K^+ -ATPase activity by incubation with omeprazole (**g–l**), inhibition of Notch signalling by incubation with the γ -secretase inhibitor DAPT (**m–r**) or downregulation of *lrd* translation by injection of an *lrd*-MO (**s–x**) resulted in reverse (leftward) heart looping (**g**, **m** and **s**), ectopic expression of *spaw* (**h** and **t**) in the right LPM, or absence of *spaw*

expression in the left LPM (**n** and **t**), and ectopic expression of *pitx2* in the right LPM (**i**, **o** and **u**). Expression of *lrd* in KV was not affected by any of these treatments (**j**, **p** and **v**). Cilia in KV seemed normal in size and distribution in embryos treated with omeprazole (**k**) or DAPT (**q**) but were severely distorted in *lrd* morphants (**w**). Leftward flow in KV of omeprazole-treated embryos and DAPT-treated embryos was normal (**l**, **r**) but was absent or very slow in *lrd* morphants (**x**). All embryo views are posterior, dorsal at the top, except **a**, **g**, **m** and **s**, which are ventral, anterior at the top, and **b**, **c**, **h**, **i**, **n**, **o**, **t** and **u**, which are dorsal, anterior at the top. Black and red arrowheads point to normal and ectopic gene expression domains, respectively. In **f**, **l**, **r** and **x** two superimposed photographs were taken 15 s apart and the second photograph was pseudocoloured in red.

control embryos (Fig. 1e; see also refs 23, 26), those of *lrd* morphants seemed fewer and shorter and were distributed in a patchy fashion in KV, which often seemed smaller and distorted (80% ($n = 10$); Fig. 1w). We also investigated the existence of a directional flow of extracellular fluid by tracking the movement of fluorescein-labelled latex beads injected inside KV. In control embryos we detected a movement of some of the injected beads towards the left side of KV in most cases (90% ($n = 10$); Fig. 1f). In contrast, beads injected into *lrd* morphants failed to move or did so at a much slower pace (90% ($n = 10$); Fig. 1x). These results indicate the existence of a directional leftward net flow inside KV that depends on *lrd* function and is associated with a normal distribution and/or length of cilia that line this structure. Our findings therefore provide strong experimental evidence in support of the 'nodal flow hypothesis' in vertebrate embryos other than the mouse.

To investigate the placement of *lrd* function in the cascade of laterality information in the zebrafish embryo, we injected *lrd*-MO into 1k-cell-stage embryos. This treatment induced LR patterning defects similar to those of embryos injected at the 1-cell stage (Table 1), showing that *Lrd* functions downstream of H^+/K^+ -ATPase activity during LR patterning. Furthermore, because MO injections at the 1k-cell stage preferentially target cells fated to form KV²⁶, these results argue against possible earlier functions of *Lrd* during LR patterning²⁴ of zebrafish embryos. Downregulation of H^+/K^+ -ATPase or Notch activities did not result in changes in *lrd* expression (Fig. 1j, p), cilia distribution or size (Fig. 1k, q) or fluid flow (Fig. 1l, r) in KV. Taken together, our results so far identify a LR information cascade in the zebrafish embryo that involves early epigenetic mechanisms, conserved in *Xenopus* and chick, that are amplified in KV by a cilia-based mechanism conserved in mouse.

Bilaterally symmetric somites are not a default state

In the course of our characterization of the laterality cascade in

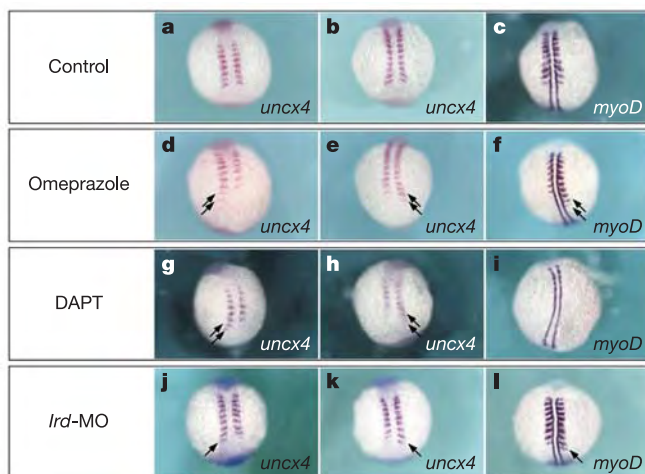


Figure 2 LR asymmetric somitogenesis in zebrafish embryos. **a–c**, Control zebrafish embryos have a symmetric number of somites at any given developmental stage, as evaluated by *uncx4* (**a, b**) and *myoD* (**c**) expression. **d–f**, Inhibition of H^+/K^+ -ATPase activity by incubation with omeprazole (**d–f**), inhibition of Notch signalling by incubation with the γ -secretase inhibitor DAPT (**g–i**) or downregulation of *lrd* translation by injection of an *lrd*-MO (**j–l**) resulted in an uneven number of somites on the left and right sides, as evaluated by *uncx4* expression at the 8-somite stage (**d, g** and **j**) and at the 10-somite stage (**e, h** and **k**). The expression of *myoD* revealed asymmetric somitogenesis at the 10-somite stage in embryos incubated with omeprazole (**f**) and *lrd* morphants (**l**), and defects in somite maturation in embryos incubated with DAPT (**i**). All embryo views are dorsal, anterior at the top. Extra somites are indicated by arrows.

zebrafish embryos, we observed that experimental manipulations that blocked the flow of LR information resulted in evident but transient LR asymmetries in somite formation. This phenotype indicated that the normal bilaterally symmetric progression of somitogenesis was somehow under the control of the LR information cascade. Thus, whereas somite formation and segmentation proceeded in a bilaterally symmetric fashion in control embryos²⁷ (Fig. 2a–c), downregulation of H^+/K^+ -ATPase activity from the 1–2-cell stage up to the bud stages (but not from mid-epiboly onwards; not shown) frequently resulted in uneven numbers of somites on the left and right sides when analysed between the 6-somite and 13-somite stages (36% ($n = 89$); Fig. 2d–f). This asymmetry was not obviously biased in any direction: 41% of the abnormal embryos had one or more somites on the left side and 59% had one or more somites on the right side. Similar results were obtained by blocking the LR information flow at the level of Notch signalling. Consistent with the crucial role of Notch signalling in the molecular clock that regulates somitogenesis²⁸, and with phenotypes of zebrafish mutants in Notch pathway components^{29–32}, incubation with DAPT from bud to early somitogenesis stages resulted in evident alterations in somite segmentation and maturation (Fig. 2i; see also ref. 21). This treatment also induced LR asymmetries in the number of somites (26% ($n = 55$); Fig. 2g, h) that were not biased towards either side of the embryo: 46% of the abnormal embryos had one or more somites on the left side and 54% had one or more somites on the right side. These results indicate that the bilaterally symmetric somitogenesis of zebrafish embryos is not a default state but rather that it requires an active mechanism of LR clock coordination, under the control of the LR information cascade.

The timing of the appearance of somite LR asymmetries after the downregulation of H^+/K^+ -ATPase or Notch activities (between the

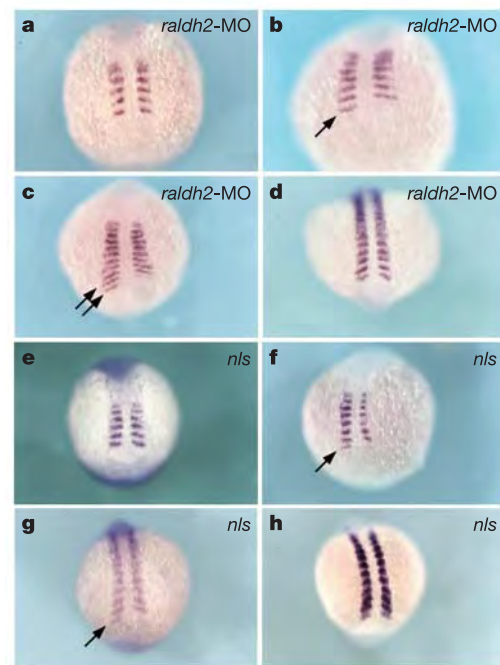


Figure 3 RA signalling coordinates LR somitogenesis in zebrafish embryos. **a–h**, *In situ* hybridization for *uncx4* expression on *raldh2* morphants (**a–d**) or *nls* mutant embryos (**e–h**), in which RA signalling is downregulated, reveals bilaterally symmetric somite numbers at the 5-somite (**a**), 6-somite (**e**) and 14-somite (**d** and **h**) stages, and increased number of somites on the left side at the 7-somite (**b**), 8-somite (**f**) and 11-somite (**c** and **g**) stages. All embryo views are dorsal, anterior at the top. Extra somites are indicated by arrows.

6-somite and 13-somite stages) is well correlated with the proposed developmental window for the role of KV in LR patterning²⁶. We therefore next asked whether the disruption of cilia function and leftward fluid flow in KV of *lrd* morphants were associated with LR asymmetric somitogenesis. In these experiments a small but significant percentage of embryos had LR asymmetries in somite formation (14% ($n = 210$); Fig. 2j–l); these were unbiased (50% of abnormal embryos had more somites on the left side and 50% had more on the right side) and were limited to embryos between the 6-somite and 13-somite stages. We then analysed zebrafish *one eye pinhead* (*oep*) mutant embryos, in which the transfer of LR information to the node is thought to occur normally but in which there is an impaired relay of laterality information from the node to the LPM³³. We found that, although *oep* mutant embryos had alterations of LR visceral asymmetries (not shown; see also ref. 33), somitogenesis proceeded in a bilaterally symmetric fashion indistinguishable from wild-type sibling embryos (not shown). Taken together, our results indicate that the cascade of LR information branches in or near KV to control the LR asymmetric patterning of internal organs and the LR symmetric segmentation of presomitic mesoderm (PSM) independently.

RA signalling coordinates LR and AP axis elongation

We next investigated the mechanism by which the LR information cascade controls the bilateral symmetry of somite formation. Nascent somites are specified in the PSM by the combined action

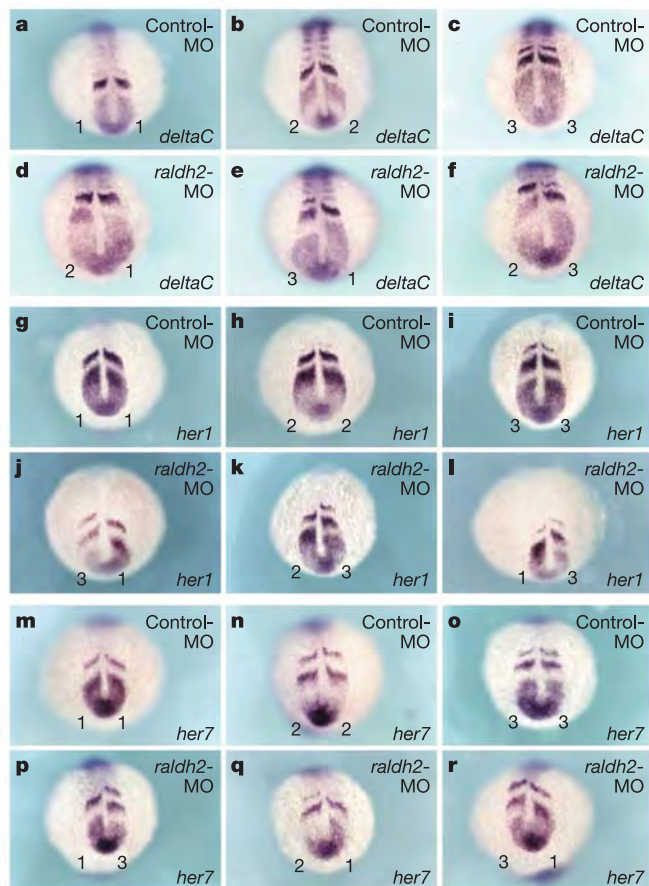


Figure 4 Desynchronization of the molecular clock in *raldh2* morphants. **a–r**, *In situ* hybridization for *deltaC* (**a–f**), *her1* (**g–l**) and *her7* (**m–r**) expression in control zebrafish embryos (**a–c**, **g–i** and **m–o**) showing synchronization between the left and right sides of the embryo, whereas *raldh2* morphants (**d–f**, **j–l** and **p–r**) have desynchronized expression patterns. All embryo views are posterior, dorsal at the top. The phase of the cycle (1, 2 or 3) is indicated under each left and right PSM.

of opposed gradients of Wnt/FGF and RA activities and by the synchronized oscillations of Notch and Wnt activities (reviewed in refs 6, 7, 28, 34). We manipulated the activity of these four pathways to ascertain their involvement in the coordination of LR somitogenesis in the zebrafish. Inhibition of FGF and/or Wnt signalling did not result in asymmetric somitogenesis in zebrafish embryos (Supplementary Fig. S2). In contrast, inhibiting Notch activity with DAPT (Fig. 2g, h) resulted in embryos with an uneven number of somites on each side, indicating that Notch activity is mechanistically involved in the coordination of LR somitogenesis. However, the fact that the observed LR asymmetries in somite formation occurred in an unbiased fashion indicates the possible

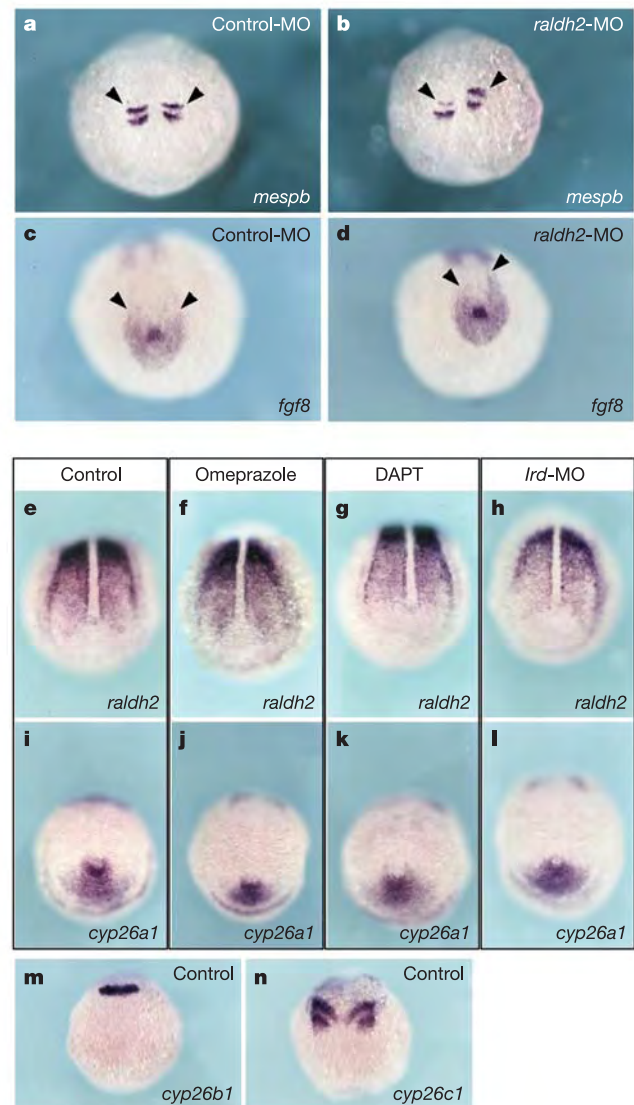


Figure 5 RA signalling counteracts the LR information flow during zebrafish somitogenesis. **a, b**, *In situ* hybridization for *mespb* in *raldh2* morphants (**b**) reveals a LR asymmetry in its expression domain compared with control embryos (**a**). **c, d**, Expression of *fgf8* in control embryos is bilaterally symmetric (**c**), whereas *raldh2* morphants show an anterior extension of the *fgf8* expression domain on the right side (**d**). **e–n**, *In situ* hybridization for *raldh2* (**e–h**) and *cyp26a1* (**i–l**) reveals complementary expression domains around the tailbud of control embryos (**e, i**), which are not altered in embryos treated with omeprazole (**f, j**) or DAPT (**g, k**) or in *lrd* morphants (**h, l**). **m, n**, The expression of *cyp26b1* and *cyp26c1* in early somitogenesis embryos is restricted to the midbrain–hindbrain boundary. All embryo views are posterior, dorsal at the top, except in **m** and **n**, which are dorsal, anterior at the top. Arrowheads point to the anterior level of gene expression.

existence of additional mechanisms that would provide the necessary laterality information to counteract the influence of the LR information flow.

Indeed, blocking endogenous RA production by using an MO against *raldh2* (the major source of RA in the context of somite formation³⁵, also known as *aldh1a2* or aldehyde dehydrogenase 1 family, member A2) resulted in a strongly biased LR asymmetric somite development. *raldh2* morphants initiated somitogenesis in a bilaterally symmetric way that was indistinguishable from that in embryos injected with control MO (Fig. 3a). However, as somite formation proceeded, a significant fraction of *raldh2* morphants developed an uneven number of somites on the left and right sides (21%, versus 0% in embryos injected with control MO ($n = 137$ and 153, respectively); Fig. 3b, c). This asymmetry was evident between the 6-somite and 13-somite stages, after which *raldh2* morphants recovered a normal bilaterally symmetric number of somites (Fig. 3d). The LR asymmetry in somitogenesis seemed strongly biased: the left side had more somites than the right one in 76% of the embryos with asymmetric somite formation. The specificity of the phenotypes induced by injecting *raldh2*-MO was further verified with zebrafish *neckless* (*nls*) mutant embryos, which lack RALDH2 activity³⁶, with similar results (Fig. 3e–h and Supplementary Table S1).

We next investigated the integrity of the LR information cascade in embryos injected with *raldh2*-MO. No alterations in the left-sided expression patterns of *spaw* or *pitx2* in the LPM were detected in *raldh2* morphants ($n = 36$ and 48, respectively; not shown), which is consistent with previous results from quail³⁷ and mouse³⁸ embryos. In contrast, we found that the left-biased asymmetries in somitogenesis caused by the downregulation of RA signalling required the existence of a normal flow of LR information. The inhibition of H^+/K^+ -ATPase activity or *lrd* function in *raldh2* morphants therefore prevented the left-sided bias of asymmetric somitogenesis that is characteristic of *raldh2* morphants (Supplementary Table S1).

Molecular basis of LR asymmetric somitogenesis

The periodicity of somite formation is thought to depend on the action of an oscillator or clock that sets the pace of segmentation in the cells of the PSM. Several components of the Notch signalling

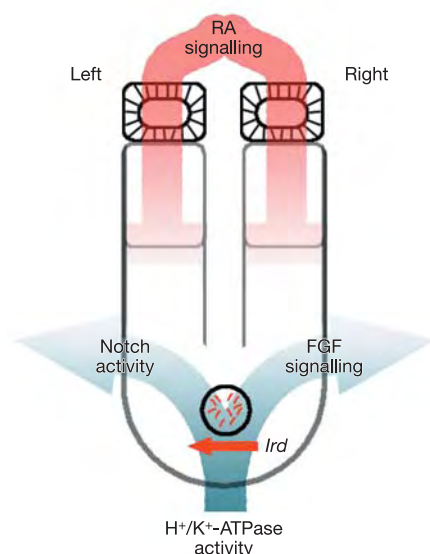


Figure 6 Crosstalk between LR and AP axes. Diagram illustrating the interactions between the mechanisms controlling LR patterning and somite formation during early somitogenesis stages, and the proposed role of RA signalling as a LR buffer of the laterality information flow (blue arrows) in the zebrafish embryo. See the text for details.

pathway have been identified as molecular readouts of this segmentation clock and have been shown to have oscillatory expression patterns (cyclic genes; reviewed in refs 39, 40). We investigated the integrity of the segmentation clock in zebrafish *raldh2*-morphant or *nls* mutant embryos by analysing the expression of the cyclic genes *deltaC*³⁰, *her1* (refs 29, 41) and *her7* (refs 42, 43). Whereas no differences were observed in the cycle phase⁴⁴ of either transcript between left and right PMS in control embryos at any developmental stage (Fig. 4a–c, g–i, m–o), clear desynchronization was evident in embryos in which RA signalling was inhibited. Thus, most *raldh2*-morphant and *nls* mutant embryos analysed between the 4-somite and 12-somite stages had LR asymmetric expression patterns of *deltaC* (25% ($n = 281$); Fig. 4d–f, and not shown), *her1* (28% ($n = 258$); Fig. 4j–l, and not shown) and *her7* (30% ($n = 147$); Fig. 4p–r, and not shown). The degree of asymmetry between the left and right PSM was variable between embryos from the same injection experiment, some having an offset of the cyclic gene expression by one phase (Fig. 4d, f, k, q) and others by two phases (Fig. 4e, j, l, p, r); still others were in the same phase of consecutive cycles (not shown). These results indicate that RA signalling is necessary for the synchronization of the molecular clock between the left and right PSM.

The observation that extra somites in *raldh2* morphants or *nls* mutant embryos are consistently the last ones formed is difficult to explain by a simple LR clock desynchronization mechanism. Instead, such a phenotype would be compatible with an anterior shift of the right determination front⁶ or a posterior shift of the left one. To investigate these possibilities we analysed the expression of *mespb*⁴¹ (one of the zebrafish orthologues of the mouse *Mesp2* segmentation polarity gene⁴⁵) after downregulation of RA signalling. In control embryos *mespb* is expressed in two or three bilaterally symmetric stripes that localize to the anterior half of the forming somite and the most anterior presumptive somite(s) (Fig. 5a; see also ref. 41). In contrast, *raldh2* morphants had striking LR asymmetries in the expression pattern of *mespb*, which most frequently appeared shifted anteriorly on the right side of the embryo (21% ($n = 35$); Fig. 5b).

Anterior shifts in the expression pattern of *Thylacine1* (*Thy1*), the *Xenopus* orthologue of *mespb*, have been reported after downregulation of RA signalling in frog embryos⁴⁶. Although a possible LR asymmetry in such shifts was not analysed in those experiments (one side of the embryo served as control), the authors showed that the AP level of *Thy1* expression was determined by the antagonism of RA and FGF signalling pathways⁴⁶. To test whether a similar situation occurred in *raldh2*-morphant zebrafish embryos, we analysed the expression of *fgf8*, which is normally expressed in a broad domain in the posterior PSM (Fig. 5c; see also ref. 47). We detected an anterior shift in the expression domain of *fgf8* on the right side of *raldh2* morphants (25% ($n = 32$); Fig. 5d).

Finally, we investigated the mechanism by which RA signalling becomes lateralized in response to the LR information cascade. One possibility is at the level of expression of enzymes responsible for the synthesis or degradation of RA or its receptors. To address this, we performed *in situ* hybridization analyses of zebrafish *raldh2* (which our results indicate encodes the RA-synthesizing enzyme relevant for preventing LR asymmetric somitogenesis), *cyp26a1*, *cyp26b1* and *cyp26c1* (encoding RA-catabolizing enzymes), and *rara2a*, *rara2b*, *rary*, *rxra*, *rxrβ*, *rxrγ* and *rxrδ* (encoding RA receptors) in control embryos of 3-somite and 6–8-somite stages, and after experimental manipulations of the LR information cascade. We could not detect obvious LR asymmetries in the expression of these transcripts at any stage analysed in control or manipulated embryos (Fig. 5e–n, Supplementary Fig. S3, and data not shown), indicating that the control of RA signalling by the LR information cascade is likely to be post-transcriptional. In this respect, a possibility that warrants further investigation is the participation of a mechanism recently uncovered in the mouse embryo node that

results in the sided accumulation of RA itself (N. Hirokawa, personal communication).

Taken together, our results demonstrate that the bilateral progression of somitogenesis in zebrafish embryos depends on RA signalling and is tightly linked to the cascade of LR organ asymmetry information. The findings presented in the accompanying paper indicate that this mechanism is conserved in mouse and chick embryos (ref. 48; P. Dollé, personal communication). Specifically, our results are consistent with a model in which the LR information flow causes transient side-restricted changes in the activities of the Notch, FGF and/or Wnt signalling pathways, which are subsequently buffered by the antagonistic action of RA signalling (Fig. 6). In the chick, the normal LR information flow induces transcription and subtle lateralization in the expression of *Fgf8* (ref. 49), *Wnt8C*⁵⁰ and the Notch pathway components *Delta-like1* and *Lunatic fringe*¹⁷, in the perinodal region. Even though comparable LR differences in the normal expression of such transcripts are not observed in zebrafish embryos, it is possible that they are too subtle to be detected by *in situ* hybridization. Alternatively, LR differences in the activity of Notch, FGF and/or Wnt signalling pathways might take place at post-transcriptional levels. In any case, our experiments show that blocking the arrival of LR asymmetric inputs to the node results in an unbiased desynchronization of somite formation between the left and right sides of the embryo. This phenotype is consistent with the existence of LR asymmetric antagonistic influences on somitogenesis on either side of the zebrafish node. In our model, the relief of RA-signalling-mediated buffering of these LR asymmetric influences, induced in *raldh2* morphants and *nls* mutant embryos, would result in their amplification, leading to a desynchronization of the Notch-based molecular clock and to an extension of the *fgf8* expression domain on the right side of the embryo. These molecular alterations would in turn result in a delayed maturation of presomitic cells in the right PSM, as compared with the left PSM, and LR asymmetric somite segmentation. Our findings identify a crosstalk between the molecular mechanisms that regulate LR patterning and AP axis extension to ensure that both processes take place in a coordinated manner. □

Methods

Zebrafish strains

Wild-type (AB), *mlc2a*-eGFP, *nls*¹²⁶ and *oep*¹²⁵⁷ zebrafish strains were used in this study. References for these strains are provided in Supplementary Methods.

Pharmacological treatments

Zebrafish embryos were incubated with omeprazole (Sigma), SCH 28080 (Sigma), DAPT (γ-secretase inhibitor IX; Calbiochem) and/or SU5402 (Calbiochem) as detailed in Supplementary Methods.

Morpholino and messenger RNA injections

Capped mRNAs encoding axin or GFP were synthesized and used as detailed in Supplementary Methods. MOs against *raldh2*, *fgf8* or *lrd*, and a control MO, were designed and used as detailed in Supplementary Methods.

In situ hybridization

Details of the whole-mount *in situ* hybridization protocol and probes used in this study are given in Supplementary Methods.

Other methods

Details of immunofluorescence analyses and visualization of fluid flow in KV are provided in Supplementary Methods.

Received 17 December 2004; accepted 1 March 2005; doi:10.1038/nature03512.

- Levin, M. Left-right asymmetry in embryonic development: a comprehensive review. *Mech. Dev.* **122**, 3–25 (2005).
- Raya, A. & Izpisua Belmonte, J. C. Sequential transfer of left-right information during vertebrate embryo development. *Curr. Opin. Genet. Dev.* **14**, 575–581 (2004).
- Bisgrove, B. W., Morelli, S. H. & Yost, H. J. Genetics of human laterality disorders: Insights from vertebrate model systems. *Annu. Rev. Genomics Hum. Genet.* **4**, 1–32 (2003).
- Hamada, H., Meno, C., Watanabe, D. & Saijoh, Y. Establishment of vertebrate left-right asymmetry. *Nature Rev. Genet.* **3**, 103–113 (2002).
- Capdevila, J., Vogan, K. J., Tabin, C. J. & Izpisua Belmonte, J. C. Mechanisms of left-right determination in vertebrates. *Cell* **101**, 9–21 (2000).

- Dubrule, J. & Pourquie, O. Coupling segmentation to axis formation. *Development* **131**, 5783–5793 (2004).
- Aulehla, A. & Herrmann, B. G. Segmentation in vertebrates: clock and gradient finally joined. *Genes Dev.* **18**, 2060–2067 (2004).
- Weinmaster, G. & Kintner, C. Modulation of Notch signaling during somitogenesis. *Annu. Rev. Cell Dev. Biol.* **19**, 367–395 (2003).
- Saga, Y. & Takeda, H. The making of the somite: molecular events in vertebrate segmentation. *Nature Rev. Genet.* **2**, 835–845 (2001).
- Zakany, J., Kmita, M., Alarcon, P., de la Pompa, J. L. & Duboule, D. Localized and transient transcription of *Hox* genes suggests a link between patterning and the segmentation clock. *Cell* **106**, 207–217 (2001).
- Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio)* (Univ. of Oregon Press, Eugene, 2000).
- Campione, M. *et al.* The homeobox gene *Pitx2*: mediator of asymmetric left-right signaling in vertebrate heart and gut looping. *Development* **126**, 1225–1234 (1999).
- Tsukui, T. *et al.* Multiple left-right asymmetry defects in *Shh*^{−/−} mutant mice unveil a convergence of the *Shh* and retinoic acid pathways in the control of *Lefty-1*. *Proc. Natl Acad. Sci. USA* **96**, 11376–11381 (1999).
- Long, S., Ahmad, N. & Rebagliati, M. The zebrafish *nodal*-related gene *southpaw* is required for visceral and diencephalic left-right asymmetry. *Development* **130**, 2303–2316 (2003).
- Kane, D. A. & Kimmel, C. B. The zebrafish midblastula transition. *Development* **119**, 447–456 (1993).
- Levin, M., Thorlin, T., Robinson, K., Nogi, T. & Mercola, M. Asymmetries in H⁺/K⁺-ATPase and cell membrane potentials comprise a very early step in left-right patterning. *Cell* **111**, 77–89 (2002).
- Raya, A. *et al.* Notch activity acts as a sensor for extracellular calcium during vertebrate left-right determination. *Nature* **427**, 121–128 (2004).
- Krebs, L. T. *et al.* Notch signaling regulates left-right asymmetry determination by inducing *Nodal* expression. *Genes Dev.* **17**, 1207–1212 (2003).
- Raya, A. *et al.* Notch activity induces *Nodal* expression and mediates the establishment of left-right asymmetry in vertebrate embryos. *Genes Dev.* **17**, 1213–1218 (2003).
- Dovey, H. F. *et al.* Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain. *J. Neurochem.* **76**, 173–181 (2001).
- Geling, A., Steiner, H., Willem, M., Bally-Cuif, L. & Haass, C. A gamma-secretase inhibitor blocks Notch signaling *in vivo* and causes a severe neurogenic phenotype in zebrafish. *EMBO Rep.* **3**, 688–694 (2002).
- Nonaka, S. *et al.* Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. *Cell* **95**, 829–837 (1998).
- Essner, J. J. *et al.* Conserved function for embryonic nodal cilia. *Nature* **418**, 37–38 (2002).
- Levin, M. Motor protein control of ion flux is an early step in embryonic left-right asymmetry. *BioEssays* **25**, 1002–1010 (2003).
- Okada, Y. *et al.* Abnormal nodal flow precedes situs inversus in *iv* and *inv* mice. *Mol. Cell* **4**, 459–468 (1999).
- Amack, J. D. & Yost, H. J. The T box transcription factor no tail in ciliated cells controls zebrafish left-right asymmetry. *Curr. Biol.* **14**, 685–690 (2004).
- Stickney, H. L., Barresi, M. J. & Devoto, S. H. Somite development in zebrafish. *Dev. Dyn.* **219**, 287–303 (2000).
- Giudicelli, F. & Lewis, J. The vertebrate segmentation clock. *Curr. Opin. Genet. Dev.* **14**, 407–414 (2004).
- Holley, S. A., Geisler, R. & Nusslein-Volhard, C. Control of *her1* expression during zebrafish somitogenesis by a *Delta*-dependent oscillator and an independent wave-front activity. *Genes Dev.* **14**, 1678–1690 (2000).
- Jiang, Y. J. *et al.* Notch signalling and the synchronization of the somite segmentation clock. *Nature* **408**, 475–479 (2000).
- Holley, S. A., Julich, D., Rauch, G. J., Geisler, R. & Nusslein-Volhard, C. *her1* and the *notch* pathway function within the oscillator mechanism that regulates zebrafish somitogenesis. *Development* **129**, 1175–1183 (2002).
- Itoh, M. *et al.* Mind bomb is a ubiquitin ligase that is essential for efficient activation of Notch signaling by *Delta*. *Dev. Cell* **4**, 67–82 (2003).
- Yan, Y. T. *et al.* Conserved requirement for *EGF-CFC* genes in vertebrate left-right axis formation. *Genes Dev.* **13**, 2527–2537 (1999).
- Bessho, Y. & Kageyama, R. Oscillations, clocks and segmentation. *Curr. Opin. Genet. Dev.* **13**, 379–384 (2003).
- Niederreither, K., Subbarayan, V., Dolle, P. & Chambon, P. Embryonic retinoic acid synthesis is essential for early mouse post-implantation development. *Nature Genet.* **21**, 444–448 (1999).
- Begemann, G., Schilling, T. F., Rauch, G. J., Geisler, R. & Ingham, P. W. The zebrafish *neckless* mutation reveals a requirement for *raldh2* in mesodermal signals that pattern the hindbrain. *Development* **128**, 3081–3094 (2001).
- Zile, M. H. *et al.* Retinoid signaling is required to complete the vertebrate cardiac left/right asymmetry pathway. *Dev. Biol.* **223**, 323–338 (2000).
- Niederreither, K. *et al.* Embryonic retinoic acid synthesis is essential for heart morphogenesis in the mouse. *Development* **128**, 1019–1031 (2001).
- Rida, P. C., Le Minh, N. & Jiang, Y. J. A Notch feeling of somite segmentation and beyond. *Dev. Biol.* **265**, 2–22 (2004).
- Pourquie, O. The segmentation clock: converting embryonic time into spatial pattern. *Science* **301**, 328–330 (2003).
- Sawada, A. *et al.* Zebrafish *Mesp* family genes, *mesp-a* and *mesp-b* are segmentally expressed in the presomitic mesoderm, and *Mesp-b* confers the anterior identity to the developing somites. *Development* **127**, 1691–1702 (2000).
- Henry, C. A. *et al.* Two linked *hairy/Enhancer of split*-related zebrafish genes, *her1* and *her7*, function together to refine alternating somite boundaries. *Development* **129**, 3693–3704 (2002).
- Oates, A. C. & Ho, R. K. *Hairy/E(spl)-related (Her)* genes are central components of the segmentation oscillator and display redundancy with the *Delta/Notch* signaling pathway in the formation of anterior segmental boundaries in the zebrafish. *Development* **129**, 2929–2946 (2002).
- Pourquie, O. & Tam, P. P. A nomenclature for prospective somites and phases of cyclic gene expression in the presomitic mesoderm. *Dev. Cell* **1**, 619–620 (2001).

45. Saga, Y., Hata, N., Koseki, H. & Taketo, M. M. *Mesp2*: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation. *Genes Dev.* **11**, 1827–1839 (1997).
46. Moreno, T. A. & Kintner, C. Regulation of segmental patterning by retinoic acid signaling during *Xenopus* somitogenesis. *Dev. Cell* **6**, 205–218 (2004).
47. Reifers, F. *et al.* *Fgf8* is mutated in zebrafish *acerebellar* (*ace*) mutants and is required for maintenance of midbrain-hindbrain boundary development and somitogenesis. *Development* **125**, 2381–2395 (1998).
48. Vermot, J. & Pourquié, O. Retinoic acid coordinates somitogenesis and left–right patterning in vertebrate embryos. *Nature* doi:10.1038/nature03488 (this issue).
49. Boettger, T., Wittler, L. & Kessel, M. FGF8 functions in the specification of the right body side of the chick. *Curr. Biol.* **9**, 277–280 (1999).
50. Rodriguez-Esteban, C., Capdevila, J., Kawakami, Y. & Izpisua Belmonte, J. C. Wnt signaling and PKA control *Nodal* expression and left-right determination in the chick embryo. *Development* **128**, 3189–3195 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Tsukui, H. Takeda and A. Smolka for sharing reagents; N. Hirokawa and P. Dollé for sharing results before publication; C. Kintner for helpful suggestions; all members of the laboratory for discussions; I. Dubova for help with fish procedures; C. Callol, T. Chapman, H. Kawakami and M. Sugii for technical assistance; and M.-F. Schwarz for help in the preparation of this manuscript. A.R. and C.R.-E. are partly supported by postdoctoral fellowships from Fundación Inbiomed, Spain. This study was funded by the NIH, the Human Frontier Science Program, and the G. Harold and Leila Y. Mathers Charitable Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.C.I.B. (belmonte@salk.edu). The sequences of zebrafish *uncx4* and *cyp26c1* cDNAs have been deposited in GenBank with accession numbers AY881012 and AY904031, respectively.

FGF-induced vesicular release of Sonic hedgehog and retinoic acid in leftward nodal flow is critical for left–right determination

Yosuke Tanaka, Yasushi Okada & Nobutaka Hirokawa

Department of Cell Biology and Anatomy, Graduate School of Medicine, University of Tokyo, Hongo, Tokyo 113-0033, Japan

The precise specification of left–right asymmetry is an essential process for patterning internal organs in vertebrates. In mouse embryonic development, the symmetry-breaking process in left–right determination is initiated by a leftward extraembryonic fluid flow on the surface of the ventral node. However, it is not known whether the signal transduction mechanism of this flow is chemical or mechanical. Here we show that fibroblast growth factor (FGF) signalling triggers secretion of membrane-sheathed objects 0.3–5 μm in diameter termed ‘nodal vesicular parcels’ (NVPs) that carry Sonic hedgehog and retinoic acid. These NVPs are transported leftward by the fluid flow and eventually fragment close to the left wall of the ventral node. The silencing effects of the FGF-receptor inhibitor SU5402 on NVP secretion and on a downstream rise in Ca^{2+} were sufficiently reversed by exogenous Sonic hedgehog peptide or retinoic acid, suggesting that FGF-triggered surface accumulation of cargo morphogens may be essential for launching NVPs. Thus, we propose that NVP flow is a new mode of extracellular transport that forms a left–right gradient of morphogens.

Rapidly rotating cilia generate leftward fluid flow in the ventral node of mammalian embryos. This flow is termed the nodal flow and is a critical event of symmetry breaking during the late-streak and early-somite stages^{1–5}. Our previous studies have shown that mutant molecular-motor-gene mice lacking cilia (*kif3a*^{−/−}, *kif3b*^{−/−}; refs 2 and 3) and those lacking motility of the cilia (*iv*/*iv*; ref. 4) are similarly impaired in generating nodal flow, in the laterality of expressed domains of left markers, and in the direction of heart looping. Thus the nodal flow, generated by motile cilia, was considered to be the earliest symmetry-breaking event that determines laterality of mammalian embryos. This nodal flow hypothesis of left–right determination is further supported by a clinical knowledge that immotile cilia syndrome is frequently associated with situs inversus⁶.

However, the signal transduction mechanism of nodal flow has remained unknown, because transported objects have not yet been visualized. Because cilia on the periphery of the ventral node are less motile, it has even been considered that mechanosensation of the fluid flow by those cilia would be sufficient to initiate the left-specific signalling cascades; one such cascade involves an increase in Ca^{2+} concentration on the left periphery of the node^{7–9}. Here we identify membrane-sheathed objects that we term ‘nodal vesicular parcels’, or NVPs, which carry Sonic hedgehog (SHH) and retinoic acid (RA) and are secreted and transported to the left by the nodal flow. Using pharmacological and molecular genetic perturbations, we further present cell biological data suggesting how fibroblast growth factor (FGF) signalling triggers NVP secretion, and how the turnover dynamics of NVPs contribute to the unidirectionality of transport.

FGF-induced Ca^{2+} increase on the left

FGF signalling in the early-somite stage is known to generate a posterior-to-anterior gradient that is essential for maintaining the caudal identity of presomitic mesoderm in the posterior region where the concentration of FGF8 is higher than a certain

threshold^{10,11}. Because FGF signalling also appears to be essential for left determination from the phenotype of *fgf8* gene knockout mice¹², we sought to label the distribution of FGF receptors by applying specific antibodies. Figure 1 shows the stainings of FGFR1, -2, and -3 that were similarly localized on nodal cilia and the perinodal cell surface. As some of the ciliary punctate signals colocalized with those of the molecular motor KIF3A¹³ (Fig. 1b), they may represent protein incorporated in the cargos of the intraciliary transport^{14,15} that enable a microtubule-based dynamic shuttling of essential proteins; suggesting that most of the nodal cilia would have an FGF sensory function. Indeed, most nodal cells contained high levels of phosphotyrosine residues without an apparent laterality (Fig. 1g; $n = 8/8$). These signals possibly represented the FGF signalling activity all over the node, because they could be significantly diminished by treatment with a specific inhibitor of the FGF receptor tyrosine kinase, SU5402 (ref. 16, Fig. 1h; $n = 5/8$, $P \leq 0.01$, Fisher’s exact probability test). Thus FGF proteins appear not to generate a left–right concentration gradient by themselves, but rather to trigger the signalling event, making *fgfs* master genes.

We next examined the effect on the left-determination process of blocking FGF signalling. Using a calcium-sensitive fluorophore Fluo3, we observed that static elevation of the Ca^{2+} level was initiated from the left margin of the node and laterally propagated towards the left lateral plate mesoderm through a succeeding chain of definitive endodermal cells, as reported⁹ (Fig. 2a). When FGF signalling was specifically antagonized for 1 h by a dominant-negative recombinant peptide of an extracellular domain of mouse FGF receptor 2 β (IIIc) fused with an Fc region of human IgG¹⁷ (FGFR-DN; 1 $\mu\text{g ml}^{-1}$) or by SU5402 (20 μM), Ca^{2+} elevation was significantly suppressed (Fig. 2b and c). We then assessed whether Ca^{2+} signals could be restored by supplementation of downstream morphogen candidates, SHH, RA and Indian hedgehog (IHH)^{12,18,19}. Recombinant N-terminal fragment of SHH protein (SHH-N; 2 $\mu\text{g ml}^{-1}$) in the presence of FGFR-DN or SU5402

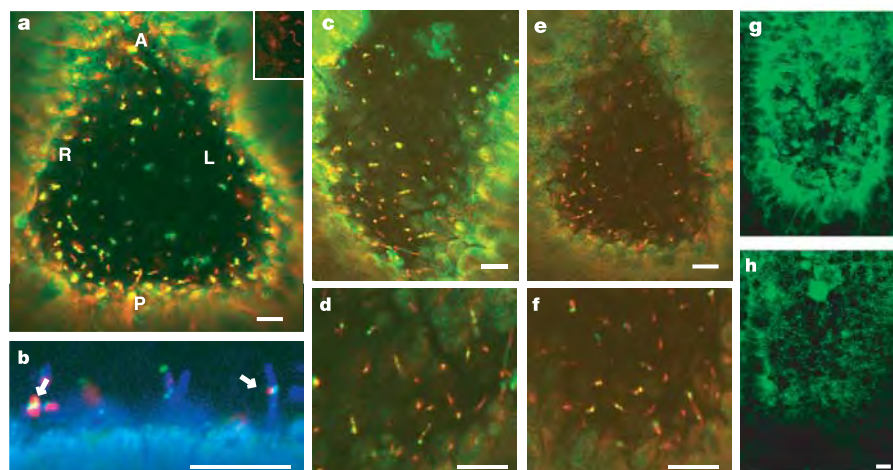


Figure 1 Expression of FGF receptors on nodal cilia. Ventral view of mouse embryos of the Crlj:CD1 (ICR) strain at 7.75 days post coitum (d.p.c.), stained against FGFR1 (**a**, **b**), FGFR2 (**c**, **d**), FGFR3 (**e**, **f**), and phosphotyrosine (**g**, **h**) in green, counterstained against acetylated tubulin (**a**, **c**–**f**, red), α -tubulin (**b**, blue), and KIF3A (**b**, red). Inset in **a**, negative

control with normal rabbit IgG (green). **h**, Reduction of phosphotyrosine by 20 μ M SU5402 treatment for 4 h. A, anterior; P, posterior; L, left; R, right. Arrows in **b**, localization of FGFR1 in intraciliary transport complex. Scale bars, 10 μ m.

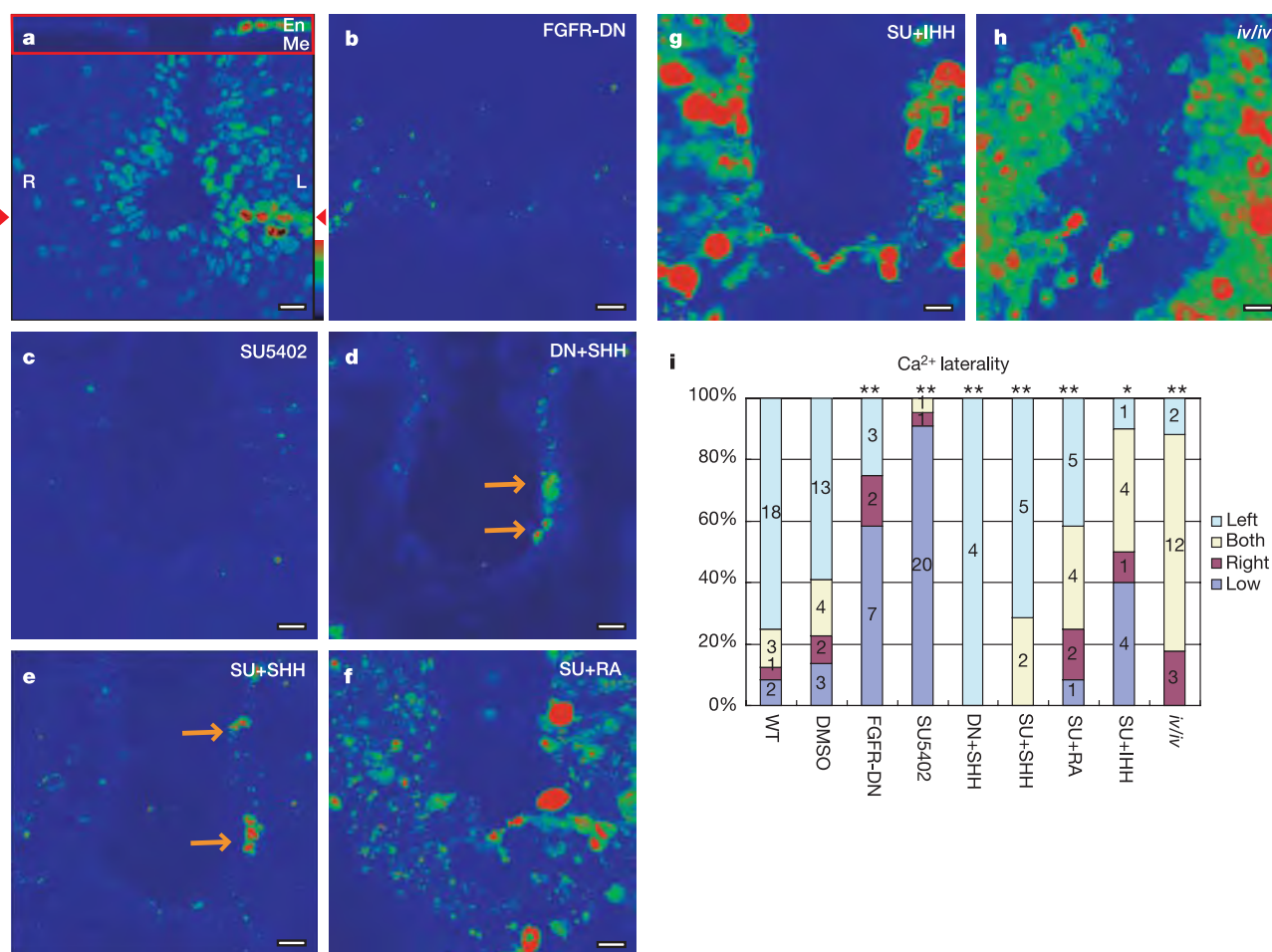


Figure 2 FGF regulates left-specific Ca^{2+} elevation. **a**–**h**, Ventral view of Ca^{2+} imaging of wild type (WT) (**a**–**g**) and *iv/iv* (**h**) embryos, pharmacologically treated as indicated, and represented by pseudocolour (signal intensity: red > green > blue > black). Upper, anterior; lower, posterior. The upper part of **a**, boxed in red, is a transverse section

reconstructed from a Z-series of images at the level between the red arrowheads (upper, ventral; lower, dorsal; En, definitive endoderm; Me, definitive mesoderm). Scale bar, 20 μ m. **i**, Statistics of laterality. *, $P < 0.03$; **, $P < 0.01$, chi-square test compared with the respective controls (WT, DMSO, SU5402).

partially restored Ca^{2+} elevation in a limited area on the left margin of the node (Fig. 2d and e). However, RA (10^{-7} M) or recombinant N-terminal fragment of IHH peptide (IHH-N; $2 \mu\text{g ml}^{-1}$) could elevate Ca^{2+} levels over a much wider area (Fig. 2f and g). These three morphogens may therefore be involved in parallel or downstream to FGF signalling in elevating Ca^{2+} levels. Interestingly, SHH and RA, but not IHH, resulted in a significant left preference for Ca^{2+} elevation (Fig. 2i), suggesting that SHH and RA are involved in an event that is central to specifying laterality, consistent with a description of their synergistic relationship in left determination¹⁸.

In the conventional two-cilia model, it is hypothesized that mechanosensory cilia on the left side specifically sense the fluid flow to elevate Ca^{2+} levels⁷. However, here we consider that the effect of the flow is more likely to be transmitted by extracellular signalling molecules, because an immotile cilium mutant *iv/iv* tended to bilaterally elevate Ca^{2+} levels (Fig. 2h and i), and that the critical parameters of the nodal flow⁴ in embryos at the 1–3 somite stage were apparently unchanged by SU5402 ($20 \mu\text{M}$) that could suppress the Ca^{2+} elevation (Fig. 3, Supplementary Movie 1). Thus, we sought to visualize the actual signalling objects being transported by the leftward nodal flow.

FGF signalling triggers NVP release

We labelled membrane lipids with a lipophilic fluorescent dye DiI solution, and observed living embryos at the 1–3 somite stage by confocal microscopy. Astonishingly, time-lapse imaging revealed that membranous parcels $0.3\text{--}5 \mu\text{m}$ in diameter were transported leftward once every 5–15 s (Fig. 4Aa; Supplementary Movie 2), which were suppressed by SU5402 or FGFR-DN (Fig. 4Ad and 4B; Supplementary Movie 3). The left side of the node tended to be brighter than the right side, suggesting that a massive transfer of lipophilic materials occurs by this transport. Scanning and transmission electron microscopy (SEM and TEM) observations (Fig. 4C and D) detected the corresponding spherical materials on the surface of nodal cells. Typically, a parcel consists of multiple lipophilic granules sheathed by an outer membrane, and is often associated with microvilli. Here we term these parcels ‘nodal vesicular parcels’ or NVPs. Time-lapse microscopy further revealed that NVPs were released one by one from dynamically protruding microvilli, flowed down the stream of the nodal flow, and were finally fragmented by ciliated surface into several small particles in proximity to the left wall; all this over approximately 30 s (Fig. 4Ab and 4Ac; Supplementary Movies 4–6). TEM observation of SU5402-treated embryos revealed a released intermediate (Fig. 4E). We also observed a significantly greater number of unsheathed particles, possibly reflecting a failure in launching (Fig. 4F; SU5402, six clumps on 43 cells; Control, two clumps on 90 cells; $P < 0.01$, chi-square test). These unique properties of release and catch would ensure the unidirectionality of the NVP flow and also enable ‘delayed activation’ of morphogens after crossing the midline, as hypothesized in our earlier model⁴.

Supplementation with SHH-N or RA again could significantly restore the NVP flow from its suppression by SU5402, but that with IHH-N could not (Fig. 4Ae, 4Af and 4B; Supplementary Movies 7–9), suggesting that SHH and RA act in the immediate downstream of FGF to launch NVPs and that IHH is committed in more downstream pathways. Their synergistic roles would explain the difference between a partial perturbation on the laterality in *Shh* knockout mice and a complete suppression of left determination in *Shh/Ihh* double-knockout mice¹⁹. This propagation towards the lateral plate mesoderm may also be facilitated by additional FGF-dependent pathways such as Notch-mediated Ca^{2+} signalling²⁰ or Cripto-mediated Nodal signalling²¹, which remain to be elucidated.

In ciliary mutants, NVPs were released but did not flow to the left (Fig. 4Ag; Supplementary Movies 10 and 11). The bilateral tendency of Ca^{2+} elevation in *iv/iv* embryos could be explained by the occasional fragmentation of NVPs caused by touching the cilia on

either side of the node. In *kif3a*^{−/−} embryos lacking cilia³, the release of NVPs was also preserved to some extent, as FGF receptors were still expressed on the cell surface (not shown). Interestingly, the numbers of NVPs that appeared in the nodes of *kif3a*^{−/−} embryos were significantly greater than those of *iv/iv* embryos (Fig. 4B). If the releasing rates are assumed to be the same, this difference possibly reflects that *kif3a*^{−/−} embryos provide a longer lifetime for NVPs. Thus, the presence of cilia appears to facilitate the fragmentation of NVPs.

NVPs carry SHH and RA to the left

Light- and electron-microscopy level immunohistochemistry using anti-SHH-N antibody 5E1 and an anti-RA antiserum revealed their association with NVPs (Fig. 5a–c, Supplementary Movie 12). Another rat anti-SHH antibody provided a similar staining pattern (not shown). Diffuse SHH signals on the cell surface were detected dominantly on the left side, indicating that SHH is transported leftward and generates a significant left–right gradient on its adsorption (lower graph of Fig. 5a).

Taken together, the above data lead to the following conclusions that SHH and RA are associated with NVPs and also facilitate the release of NVPs. We propose here a ‘shuttle-bus model’ for NVP release, wherein newly arrived passengers or the morphogens are randomly and continuously boarding on many shuttle buses or NVPs at their releasing sites, and each bus launches separately when it reaches a certain capacity. The effect of FGF signalling on the release is considered to be mainly indirect, because the release of NVPs was in a one-by-one manner rather than all at once. By supposing that FGF signalling might facilitate the local supply of morphogens from their internal source such as the Golgi apparatus, we could explain that the launch was halted by FGF-receptor

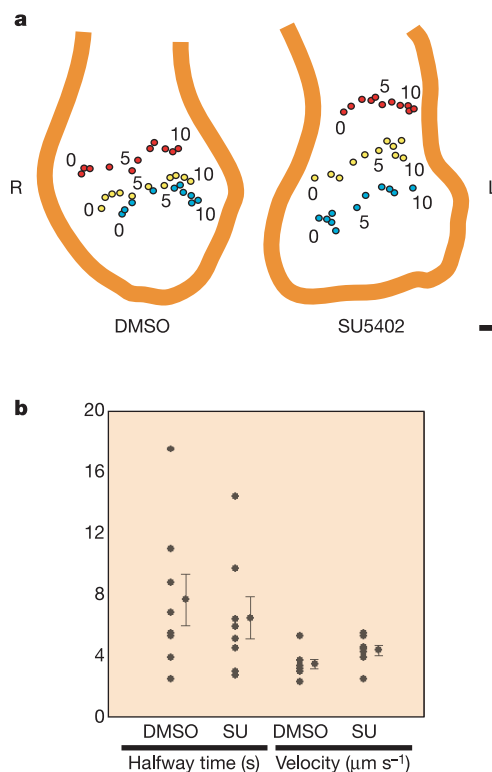


Figure 3 Leftward fluid flow is not affected by SU5402 treatment. **a**, Camera lucida of three traces of fluorescent beads for DMSO (carrier-) or SU5402-treated embryos at the 1–3 somite stage, plotted every 1 s for 10 s. Scale bar, $10 \mu\text{m}$. **b**, Statistics of **a**. Duration from the right wall to the midline (halfway time) and mean velocities of every bead are plotted. Error bars, mean $\pm$ s.e.m.

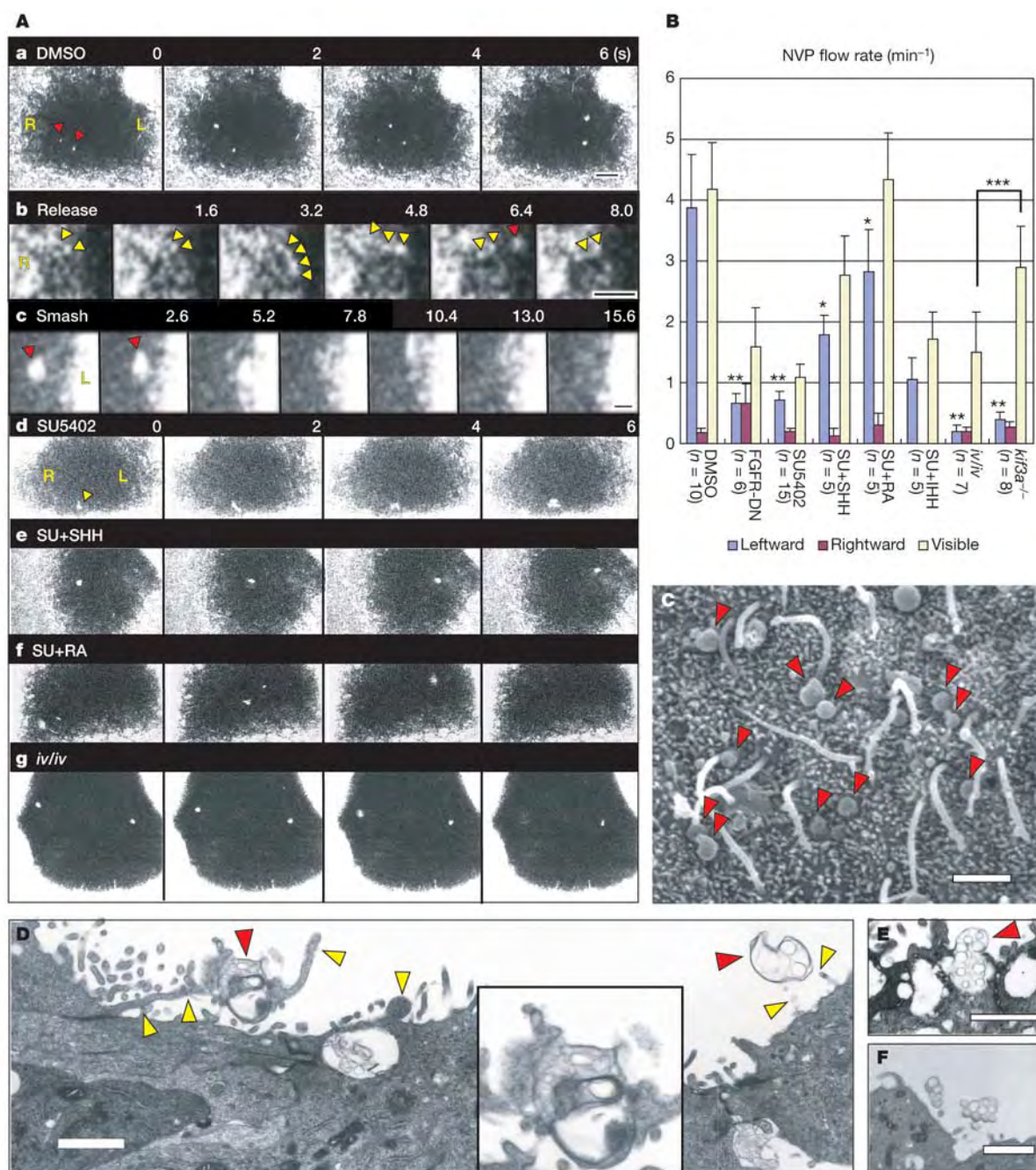


Figure 4 Nodal flow of NVPs is dependent on FGF. **A**, Time-lapse images of the ventral nodes labelled by Dil showing NVP flow. **a** and **d–g** are at the same magnification. Scale bars, 10 μm (**a**), 1 μm (**b** and **c**). Yellow and red arrowheads, respectively, show membrane protrusions and detached NVPs. **B**, Statistics of the NVPs crossing the midline and total visible NVPs per minute. Error bars, mean $\pm$ s.e.m. *, $P < 0.03$ with SU5402

(leftward); **, $P < 0.005$ with DMSO (leftward); ***, $P < 0.01$, Welch's t -test. **C**, SEM image of nodal surface. Red arrowheads, NVPs. Scale bar, 2 μm . **D–F**, TEM image of the nodal surface of control (**D**) and SU5402-treated (**E**, **F**) embryos. Red arrowheads, emerging NVPs. Yellow arrowheads, microvilli associated with NVPs. Scale bars, 1 μm . Inset in **D**, higher magnification of a typical NVP.

inhibitors and easily restored by exogenous SHH-N or RA. Because a supplementation of either morphogen was sufficient for releasing NVPs, this microvilli-dependent launching machinery may be activated synergistically by both morphogens, probably via a lipid-sensing function of Dispatched protein²². Future research should provide the molecular details of the NVP-releasing mechanism, for example in respect to the functional diversity among different FGF receptors.

Our data provide direct evidence that the nodal flow transports NVP-associated morphogens towards the left, which is probably a critical phenomenon of symmetry breaking in mammalian

embryos. At the molecular level, we clarified the cell biology of NVP flow (as summarized in Fig. 5d), which further supplies general insights for the releasing mechanism of essential morphogens. The FGF-dependent release of NVPs carrying SHH and RA changes our understanding of the relationship between FGF, SHH and RA—the morphogens positively involved in the left-determination process^{12,18,19}. These morphogens are known to act synergistically or to generate feedback loops in various phases of development^{12,23–26}, so NVP release will serve as a good model system to explore further the molecular mechanism of FGF-dependent secretion of SHH and RA.

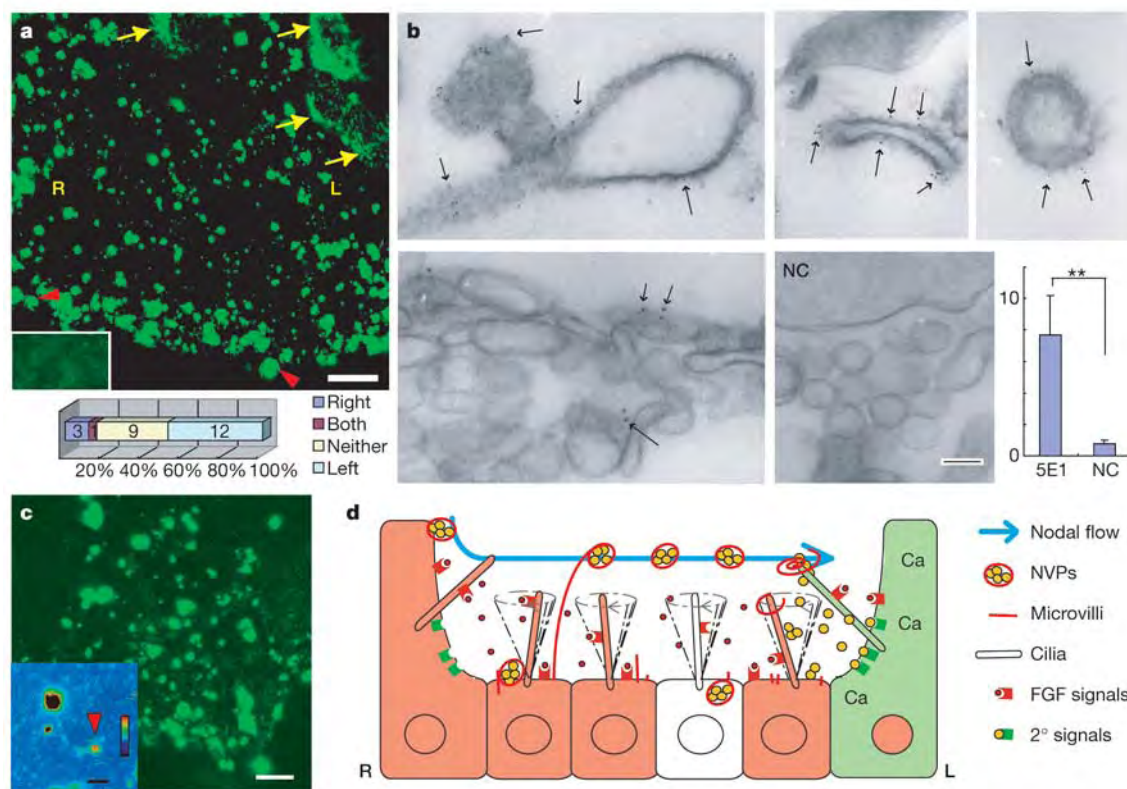


Figure 5 SHH and RA are associated with NVPs. **a**, Indirect immunofluorescence image of a wild-type node with anti-SHH antibody 5E1. Arrows, diffuse labelling on the left margin. Arrowheads, sheathed particles. Scale bar, 10 μ m. Inset in **a**, negative control with normal mouse IgG. The lower graph in **a** shows statistics of the laterality of the side diffusely labelled. **b**, Immuno-EM images with 5E1 or normal mouse IgG (NC) of NVPs. Arrows indicate gold labels. Scale bar, 100 nm. The graph shows statistical significance in the number of gold labels per NVP. **, $P < 0.01$, Welch's t -test. Error bars,

mean $\pm$ s.e.m. **c**, Indirect immunofluorescence image of a non-permeabilized embryo probed with surface RA. Scale bar, 10 μ m. Inset, pseudocoloured image of a signal (arrowhead) localized on the tip of a cellular protrusion. Scale bar, 2 μ m. **d**, Schematic representation of the NVP flow carrying SHH and RA to the left, powered by rotation of the nodal cilia. The release is triggered by FGF signalling and mediated with dynamically protruding microvilli, and the fragmentation on the left is facilitated by cilia.

Indeed, the association of SHH and RA with NVPs will shed new light on the classical problem of how the lipid-modified active form of SHH-N protein (termed SHH-Np) tethered with plasma membrane could serve in long-range signalling²⁷. The structure of an NVP—consisting of a membrane sheath and a lipid core—enables lipophilic morphogens to be transported over long distances without being solubilized in water. Besides, fragmentation of NVPs on the left side facilitates the leftward unidirectionality of NVP transport. Fragmented NVPs are scattered onto the left wall of the node, and are easily associated with the cell surface. The possible role of cilia in this fragmentation partially explains the dependency of Shh signalling on cilia that is suggested from earlier genetic evidence^{3,28}.

Finally, the microvilli-dependent releasing mechanism of NVPs suggests a new process in the extracellular release of morphogens. Although it is known that Shh signalling of the *Drosophila* imaginal disc is propagated by a microvillum-like protrusion called a cytoneme²⁹, and that FGF is essential for generating cytonemes from *Drosophila* air sac cells³⁰, the release of such large vesicular parcels from dynamically protruding microvilli at present seems unique to this system, having possibly been evolved to increase the surface area of the morphogen complex so that it can be efficiently transported by the leftward nodal flow.

Methods

Semi-whole-mount immunohistochemistry

The caudal portion of embryos was dissected in PB1 medium (Wako Pure Chemicals) containing 10% fetal bovine serum (FBS)³¹ and fixed in 4% paraformaldehyde (PFA)/phosphate-buffered saline (PBS) or 2% PFA/0.1% glutaraldehyde (GA)/PBS (for RA

staining) for 20 min at 37 °C in a 1.5-ml microtube. It was then permeabilized with 0.1% Triton X-100/PBS for 5 min at 20 °C, blocked in 10% bovine serum albumin/PBS, and the following primary antibodies and AlexaFluor-conjugated secondary antibodies (Molecular Probes) subtracted with embryo acetone powder were sequentially applied, and mounted on an APS-coated glass slide with a 0.2- μ m-thick silicon rubber spacer and a #00 coverslip (Matsunami Glass) for observation with a Zeiss LSM510 confocal laser scanning microscope. For the primary antibodies, rabbit anti-FGFR1–3 and mouse anti-acetylated tubulin clone 6-11B-1 were purchased from Sigma, mouse anti-phosphotyrosine clone PY20 from BD Transduction Laboratories, rat anti- α -tubulin clone YOL1/34 from Chemicon, rabbit anti-RA from QED Bioscience, and rat anti-SHH from R&D Systems. Monoclonal anti-KIF3A antibody was raised as previously described¹³. The 5E1 monoclonal antibody against SHH-N developed by T. M. Jessell was obtained from the Developmental Studies Hybridoma Bank (developed under the auspices of the NICHD and maintained by The University of Iowa, Department of Biological Sciences).

Live imaging

Ca²⁺ imaging and DiI labelling was performed following the manufacturers' protocols (Molecular Probes). Briefly, whole-mount embryos were loaded into PB1/FBS containing either 10 μ M fluo-3AM and 0.02% Pluronic F-127, or 4 μ M DiI, for 30 min at room temperature. The embryos were then cultured in modified DR75 medium³¹—75% IC-rat serum (Charles River Japan), 25% DMEM (Gibco-BRL), 1 mM glutamine, 100 μ M Na-pyruvate, 1 mM Hepes of pH 7.0—supplemented with or without 20 μ M SU5402 (Calbiochem) in 0.1% DMSO, recombinant peptides (R&D Systems), or 10^{−7} M all-trans retinoic acid (Sigma) for 1 h in 5% CO₂ at 37 °C, dissected in PB1/FBS, and observed in the presence of the drugs. The fluid flow was observed with fluorescent beads as previously described⁴.

Electron microscopy

Pre-embedding immunoelectron microscopy was performed basically as previously described³². Embryos were fixed in 3% PFA/0.1% GA/PBS, quenched with 1 mg ml^{−1} NaBH₄ in PBS, permeabilized with 0.1% Triton X-100 in PBS, and incubated with an anti-SHH antibody 5E1 or a normal mouse IgG. Then, an anti-mouse IgG antibody conjugated with 5 nm colloidal gold (British Biocell International) was applied, they were embedded

in Quetol-812 resin, sectioned to 1 μ m thick using a Leica ultramicrotome to check the signals, and finally re-embedded for ultrathin sectioning to be observed with a Jeol 1200-Ex transmission electron microscope. For SEM observation, embryos were prepared and observed with a Jeol JSM-5510LV scanning electron microscope as described².

Received 16 September 2004; accepted 18 February 2005; doi:10.1038/nature03494.

- Hirokawa, N. Stirring up development with the heterotrimeric kinesin KIF3. *Traffic* **1**, 29–34 (2000).
- Nonaka, S. *et al.* Randomization of left–right asymmetry due to loss of nodal cilia generating leftward flow of extra embryonic fluid in mice lacking KIF3B motor protein. *Cell* **95**, 829–837 (1998).
- Takeda, S. *et al.* Left-right asymmetry and kinesin superfamily protein KIF3A: new insights in determination of laterality and mesoderm induction by *kif3A*^{−/−} mice analysis. *J. Cell Biol.* **145**, 825–836 (1999).
- Okada, Y. *et al.* Abnormal nodal flow precedes situs inversus in *iv* and *inv* mice. *Mol. Cell* **4**, 459–468 (1999).
- Nonaka, S., Shiratori, H., Saijoh, Y. & Hamada, H. Determination of left-right patterning of the mouse embryo by artificial nodal flow. *Nature* **418**, 96–99 (2002).
- Afzelius, B. A. Cilia-related diseases. *J. Pathol.* **204**, 470–477 (2004).
- Tabin, C. J. & Vogan, K. J. A two-cilia model for vertebrate left-right axis specification. *Genes Dev.* **17**, 1–6 (2003).
- Yokoyama, T. Motor or sensor: A new aspect of primary cilia function. *Anat. Sci. Int.* **79**, 47–54 (2004).
- McGrath, J., Somlo, S., Makova, S., Tian, X. & Brueckner, M. Two populations of node monocilia initiate left-right asymmetry in the mouse. *Cell* **114**, 61–73 (2003).
- Dubrule, J., McGrew, M. J. & Pourquie, O. FGF signalling controls somite boundary position and regulates segmentation clock control of spatiotemporal *Hox* gene activation. *Cell* **106**, 219–232 (2001).
- Dubrule, J. & Pourquie, O. *fgf8* mRNA decay establishes a gradient that couples axial elongation to patterning in the vertebrate embryo. *Nature* **427**, 419–422 (2004).
- Meyers, E. N. & Martin, G. R. Differences in left-right axis pathways in mouse and chick: functions of FGF8 and SHH. *Science* **285**, 403–406 (1999).
- Kondo, S. *et al.* KIF3A is a new microtubule-based anterograde motor in the nerve axon. *J. Cell Biol.* **125**, 1095–1107 (1994).
- Rosenbaum, J. L. & Witman, G. B. Intraflagellar transport. *Nature Rev. Mol. Cell Biol.* **3**, 813–825 (2002).
- Scholey, J. M. Intraflagellar transport. *Annu. Rev. Cell Dev. Biol.* **19**, 423–443 (2003).
- Mohammadi, M. *et al.* Structures of the tyrosine kinase domain of fibroblast growth factor receptor in complex with inhibitors. *Science* **276**, 955–960 (1997).
- Kos, F. J. & Chin, C. S. Costimulation of T cell receptor-triggered IL-2 production by Jurkat T cells via fibroblast growth factor receptor 1 upon its engagement by CD56. *Immunol. Cell Biol.* **80**, 364–369 (2002).
- Tsukui, T. *et al.* Multiple left-right asymmetry defects in *Shh*^{−/−} mutant mice unveil a convergence of the Shh and retinoic acid pathways in the control of *Lefty-1*. *Proc. Natl Acad. Sci. USA* **96**, 11376–11381 (1999).
- Zhang, X. M., Ramelho-Santos, M. & McMahon, A. P. *Smoothed* mutants reveal redundant roles for Shh and Ihh signaling including regulation of L/R asymmetry by the mouse node. *Cell* **105**, 781–792 (2001).
- Raya, Á. *et al.* Notch activity acts as a sensor for extracellular calcium during vertebrate left-right determination. *Nature* **427**, 121–128 (2004).

- Mathieu, J. *et al.* Nodal and Fgf pathways interact through a positive regulatory loop and synergize to maintain mesodermal cell populations. *Development* **131**, 629–641 (2004).
- Kawakami, T. *et al.* Mouse dispatched mutants fail to distribute hedgehog proteins and are defective in hedgehog signaling. *Development* **129**, 5753–5765 (2002).
- Ye, W., Shimamura, K., Rubenstein, J. L. R., Hynes, M. A. & Rosenthal, A. FGF and Shh signals control dopaminergic and serotonergic cell fate in the anterior neural plate. *Cell* **93**, 755–766 (1998).
- Kessaris, N., Jansen, F., Rubin, L. L. & Richardson, W. D. Cooperation between sonic hedgehog and fibroblast growth factor/MAPK signalling pathways in neocortical precursors. *Development* **131**, 1289–1298 (2004).
- Scherz, P. J., Harfe, B. D., McMahon, A. P. & Tabin, C. J. The limb bud Shh-Fgf feedback loop is terminated by expansion of former ZPA cells. *Science* **305**, 396–399 (2004).
- Rice, R. *et al.* Disruption of Fgf10/Fgfr2b-coordinated epithelial-mesenchymal interactions causes cleft palate. *J. Clin. Invest.* **113**, 1692–1700 (2004).
- Zeng, X. *et al.* A freely diffusible form of Sonic hedgehog mediates long-range signalling. *Nature* **411**, 716–720 (2001).
- Huangfu, D. *et al.* Hedgehog signalling in the mouse requires intraflagellar transport proteins. *Nature* **426**, 83–87 (2003).
- Ramírez-Weber, F.-A. & Kornberg, T. B. Cytonemes: cellular processes that project to the principal signalling center in *Drosophila* imaginal discs. *Cell* **97**, 599–607 (1999).
- Sato, M. & Kornberg, T. B. FGF is an essential mitogen and chemoattractant for the air sacs of the *Drosophila* tracheal system. *Dev. Cell* **3**, 195–207 (2002).
- Sturm, K. & Tam, P. P. L. Isolation and culture of whole postimplantation embryos and germ layer derivatives. *Methods Enzymol.* **225**, 164–190 (1993).
- Tanaka, Y., Kawahata, K., Nakata, T. & Hirokawa, N. Chronological expression of microtubule-associated proteins (MAPs) in EC cell P19 after neuronal induction by retinoic acid. *Brain Res.* **596**, 269–278 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Tam and colleagues at CMRI for advice on performing embryo dissection and whole-embryo culture. We also thank S. Takeda, R. Takemura, J. Teng, Y. Noda, H. Sato, N. Onouchi, H. Fukuda, M. Sugaya, T. Akamatsu, T. Aizawa and others from the Hirokawa laboratory for providing materials, discussions and technical assistance. This study has been supported by a Center of Excellence Grant-in-Aid (to N.H.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author contributions Y.O. helped to produce Fig. 3.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.H. (hirokawa@m.u-tokyo.ac.jp).

A link between prompt optical and prompt γ -ray emission in γ -ray bursts

W. T. Vestrand¹, P. R. Wozniak¹, J. A. Wren¹, E. E. Fenimore¹, T. Sakamoto², R. R. White¹, D. Casperson¹, H. Davis¹, S. Evans¹, M. Galassi¹, K. E. McGowan¹, J. A. Schier³, J. W. Asa³, S. D. Barthelmy², J. R. Cummings², N. Gehrels², D. Hullinger², H. A. Krimm², C. B. Markwardt², K. McLean¹, D. Palmer¹, A. Parsons² & J. Tueller²

¹Los Alamos National Laboratory, Space Science and Applications Group, ISR-1, MS-D466, Los Alamos, New Mexico 87545, USA

²NASA Goddard Space Flight Center, Code 661, Greenbelt, Maryland 20771, USA

³The Pilot Group, 128 West Walnut Avenue, Unit C, Monrovia, California 91016, USA

The prompt optical emission that arrives with the γ -rays from a cosmic γ -ray burst (GRB) is a signature of the engine powering the burst, the properties of the ultra-relativistic ejecta of the explosion, and the ejecta's interactions with the surroundings^{1–5}. Until now, only GRB 990123 had been detected⁶ at optical wavelengths during the burst phase. Its prompt optical emission was variable and uncorrelated with the prompt γ -ray emission, suggesting that the optical emission was generated by a reverse shock arising from the ejecta's collision with surrounding material. Here we report prompt optical emission from GRB 041219a. It is variable and correlated with the prompt γ -rays, indicating a common origin for the optical light and the γ -rays. Within the context of the standard fireball model of GRBs, we attribute this new optical component to internal shocks driven into the burst ejecta by variations of the inner engine. The correlated optical emission is a direct probe of the jet isolated from the medium. The timing of the uncorrelated optical emission is strongly dependent on the nature of the medium.

Starting on 19 December 2004 at 01:42:18 UT, high-energy emission from a bright and very long duration γ -ray burst, named GRB 041219a, was measured by both the IBIS ('imager on board the INTEGRAL satellite') detector of the INTEGRAL satellite⁷ and the Burst Alert Telescope (BAT) of the Swift satellite⁸. The 15–350-keV fluence measured by the Swift BAT was approximately 1.55×10^{-4} erg cm⁻², placing it among the top few per cent of the 1,637 GRB events listed in the comprehensive fourth BATSE (Burst and Transient Source Experiment) catalogue⁹. The duration of γ -ray emission from GRB 041219a was approximately 520 s, making it one of the longest ever measured.

One of our RAPTOR ('rapid telescopes for optical response') telescopes¹⁰ began optical imaging of the GRB 041219a region at 01:44:13 UT, just 8 s after receipt of the INTEGRAL alert. The long duration of the burst allowed RAPTOR-S to measure the optical emission in a series of 30-s images for an unprecedented 6.4 min while prompt γ -rays were being emitted. At the location of an infrared transient identified¹¹ in subsequent images (starting at 01:49:18 UT) taken by the PAIRITEL telescope, our images show an earlier flash of optical emission (see Fig. 1) temporally coincident with the main γ -ray pulses. At its peak, the optical flash reached a measured R_c -band magnitude of $R_c = 18.6 \pm 0.1$ mag. However, the location of the event placed it in the Galactic plane and in a direction with high optical extinction (Galactic longitude and latitude: $l = 120^\circ$, $b = +0.1^\circ$). Using standard extinction maps¹², we estimate an R -band extinction of ~ 4.9 mag, but the true extinction may be larger¹³. Correcting for the nominal extinction, the peak flux we measured corresponds to a peak optical magnitude of $R_c \approx 13.7$ mag. (Error analysis and our transformation of unfiltered instrumental magnitudes to standard R_c -band

magnitudes employing standard stars¹⁴ are discussed in Supplementary Information.)

Light curves for prompt optical emission and prompt γ -ray emission from GRB 041219a are shown in Fig. 2a. The optical light curve shows: the onset of an optical flash as the dominant first γ -ray pulse begins, peak brightness during the first γ -ray pulse, continued optical emission during the secondary γ -ray peak, and a decay of the optical emission to below our detection threshold during the tertiary γ -ray enhancement.

Optical emission has been detected during the interval of prompt γ -ray emission only once before⁶, for GRB 990123. Except for an overall temporal scaling factor—GRB 041219a was about 6 times longer—the temporal morphology of the two γ -ray light curves is remarkably similar (Fig. 2). Like GRB 041219a, the γ -ray light curve for GRB 990123 had a precursor followed by a much larger primary pulse, a secondary pulse, and a smaller-amplitude tertiary flux enhancement composed of minor pulses. But in contrast to GRB 041219a, the optical light curve from GRB 990123 was low during the primary γ -ray pulse and, though more sparsely sampled, reached peak brightness after the second major pulse. This anti-correlation suggests that prompt optical emission from GRB 990123 was generated by a different process from the prompt γ -rays. The consensus interpretation is that the delayed optical peak is generated by a reverse shock^{1,3,15}, an interpretation supported by detections of the predicted rise to a peak radio flux about one day after the burst¹⁶.

For GRB 041219a, we find that the observed optical light curve is well fitted by assuming that the generation of prompt optical emission is correlated with the generation of prompt γ -ray

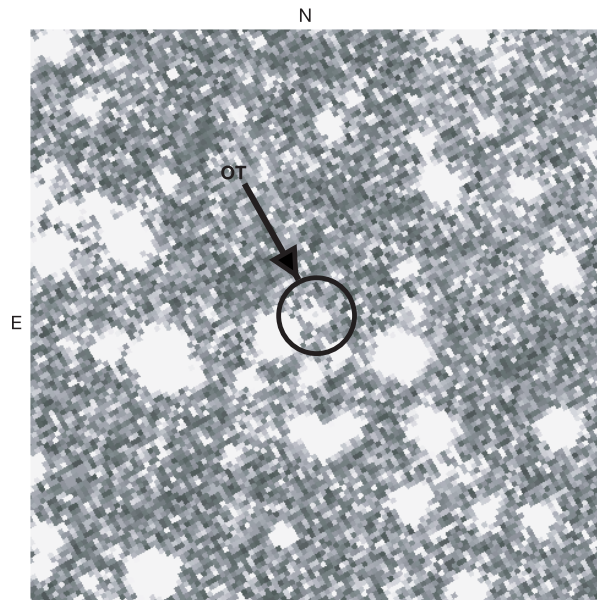


Figure 1 The prompt optical emission detected from GRB 041219a. This finder chart shows the location (right ascension 00 h 24 min 27.7 s, declination $+62^\circ 50' 33.5''$ (J2000)) of the prompt optical flash that we detected, simultaneously with the prompt γ -ray emission detected by the INTEGRAL⁷ and Swift⁸ satellites, during the time interval 01:45:41–01:49:01 UT on 2004 December 19. The location of the optical transient (OT) is identical to that found both for the subsequent infrared transient¹¹ and the optical counterpart²⁹ measured later during the late afterglow phase. Our observations of the prompt optical emission were obtained by RAPTOR-S, a 0.4-m, f/5, fully autonomous rapid response telescope owned by Los Alamos National Laboratory and located at an altitude of 2,500 m in the Jemez Mountains of New Mexico. The CCD camera employed for those observations has a $1,056 \times 1,027$ pixel format, back-illuminated, Marconi CCD47-10 chip with 13- μ m pixels.

emission. By integrating the observed 15–350-keV flux measured by Swift during the optical exposure intervals and multiplying by a derived constant optical to high-energy flux ratio, we predicted the optical light curve expected if the optical emission and γ -ray emission were perfectly correlated. As shown by the circles in Fig. 3, this simple constant-flux-ratio assumption predicts both the fast rise of the prompt optical emission observed at the start of the primary γ -ray pulse and the rapid decline observed after that dominant pulse. Our derived R_c -band optical to γ -ray flux logarithmic colour ratio for GRB 041219a is $(R_c - \gamma) = -2.5$ or, after correcting for an R-band extinction of 4.9 mag, $(R_c - \gamma) = 12.3$.

The fast rise of optical emission simultaneous with the dominant γ -ray pulse, and a general correlation with the prompt γ -ray emission would naturally arise if emission in both energy bands

was generated by a common mechanism. The broadband spectra measured during the optical observation intervals are shown in Fig. 4. Modelling of the observed spectra is beyond the scope of this Letter, but it could distinguish between emission mechanisms and provide important constraints on physical conditions in the emitting region. A particularly attractive possibility, within the standard internal–external model for GRB fireballs, is that the prompt optical emission observed in GRB 041219a is a low-energy tail of the synchrotron emission¹⁷ generated by internal shocks in the GRB outflow². In that model, a nearly constant optical to γ -ray flux ratio requires cooling times short compared to the expansion time, and therefore magnetic fields near equipartition in the ejecta. However, possibilities exist for the emission mechanism—including, for example, saturated comptonization, which can generate correlated optical and γ -ray emission¹⁸.

Internal shock models² typically predict fainter prompt optical emission than do reverse shock models. Using the γ -ray fluxes measured for GRB 990123 and scaling by 1.2×10^{-5} (from the $(R_c - \gamma)$ colour derived for GRB 041219a), we predict significantly lower optical fluxes than measured in GRB 990123—except for the first point in the optical light curve. That first optical measurement, which occurred during the dominant γ -ray pulse, is consistent, within the prediction uncertainty, with the value predicted for an internal shock using the $(R_c - \gamma)$ colour for GRB 041219a. But after

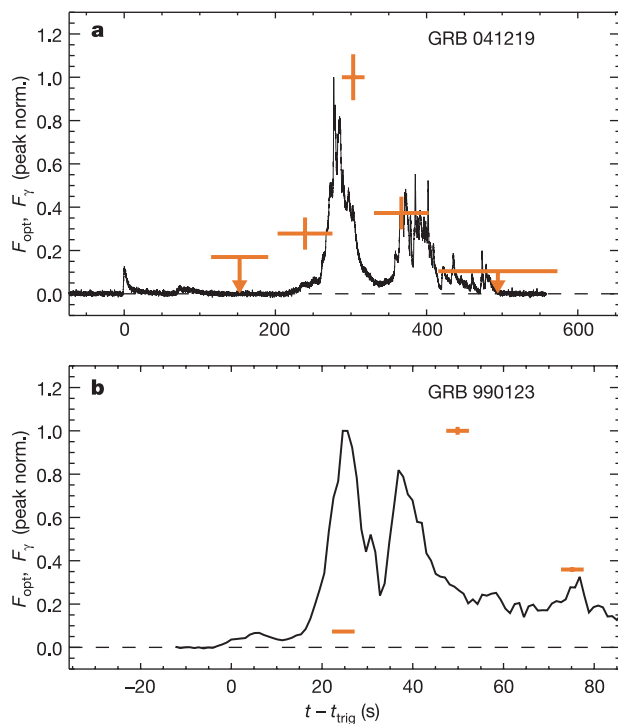


Figure 2 Comparison of the prompt γ -ray and prompt optical light curves measured⁹ for both GRB 041219a and GRB 990123. The black trace in **a** shows the relative (normalized to the peak value) γ -ray flux, F_γ , measured for GRB 041219a by the Swift satellite³⁰ as a function of elapsed time, $t - t_{\text{trig}}$, after recognition of the onset of a GRB at trigger time $t_{\text{trig}} = 01:42:18$ UT. The γ -ray light curve for GRB 041219a shows an outburst of precursor emission, followed by a mostly quiet period of 250 s, a primary pulse peaking at about 280 s, a secondary pulse centred at about 380 s, and finally a smaller-amplitude tertiary flux enhancement composed of minor pulses starting at 420 s after the trigger. The black trace in **b** shows the γ -ray light curve for GRB 990123, which is remarkably similar, except for a temporal scaling factor, to that measured for GRB 041219a. The relative optical fluxes, F_{opt} , are indicated by red crosses on both panels, with observing intervals denoted by horizontal red lines and 1σ flux error bars represented by red vertical lines. Together the panels show that the relationship between optical emission and γ -ray emission was quite different for the two events. Notice that the early optical flux from GRB 990123 was relatively low during the most intense γ -ray peak and only peaked in the optical⁶ after the two primary γ -ray peaks. For GRB 041219a, on the other hand, the prompt optical emission rises rapidly at the start and reaches a maximum during the primary γ -ray pulse, declines but persists during the secondary γ -ray pulse, and then fades below the detection threshold during the tertiary γ -ray enhancement. At peak brightness, GRB 041219a reached $R_c = 18.6 \pm 0.1$ mag, corresponding to an estimated peak magnitude of $R_c \approx 13.7$ after correction for extinction by dust.

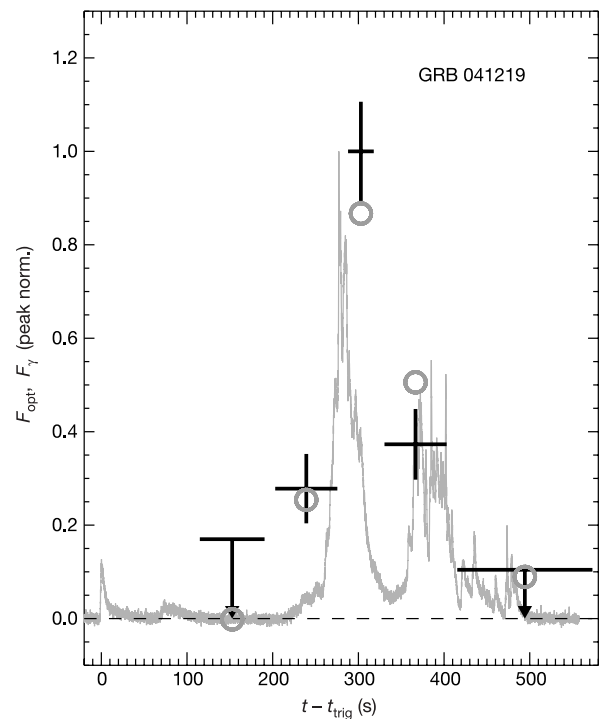


Figure 3 The measured optical light curve and that predicted for GRB 041219a assuming a constant prompt optical to prompt γ -ray flux ratio. All the optical photometry measurements are derived from stacks of two 30-s images separated by an eight-second readout time, except during the dominant γ -ray peak where a single image yielded a $>9\sigma$ detection. The γ -ray fluxes used for this comparison are derived by integrating the 15–350-keV counting rate measured by the Swift BAT, plotted as the grey trace, during the optical observation intervals. The black crosses show the actual measurements, and the circles show the predicted values. The error bars for detections are given as 1σ values, and non-detections are plotted as 2σ upper limits. The reduced χ^2 for the best-fitting model for GRB 041219a, with $F_{\text{opt}}/F_\gamma = 1.3 \times 10^{-7}$ (1.2×10^{-5} after correcting for extinction), is $\chi^2/\text{d.f.} = 1.79$ (4 degrees of freedom, d.f.). In contrast, the best-fitting model employing a constant flux ratio to predict the GRB 990123 optical light curve yields a reduced χ^2 of $\chi^2/\text{d.f.} = 1,950.65$ (2 d.f.).

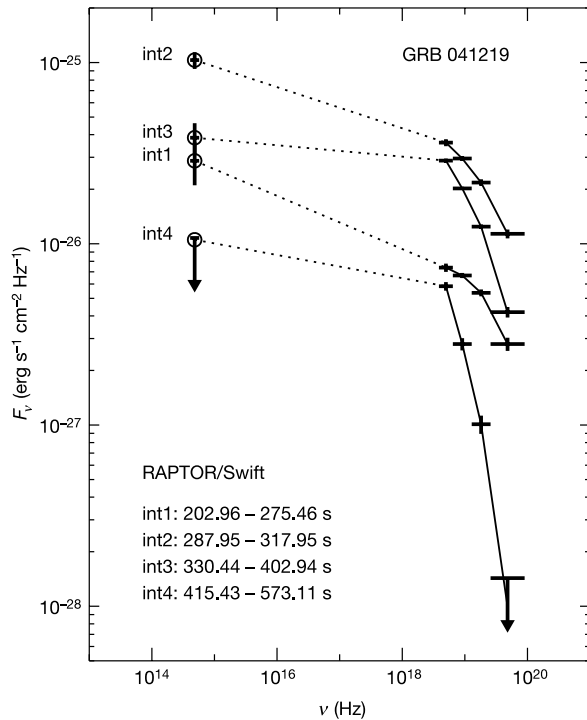


Figure 4 Broad-band spectra of GRB 041219a, here plotted in flux density F_ν , as a function of observed frequency ν , measured during the period of simultaneous prompt optical and γ -ray emission. The RAPTOR-S optical measurements, after correcting for a nominal 4.9 magnitudes of extinction, are shown as circles. Simultaneous high-energy measurements from the BAT instrument on board the Swift satellite³⁰ are shown as crosses. All the error bars represent the 1σ statistical errors. We estimate that the systematic uncertainty for the normalization of the optical fluxes is about a factor of three due to uncertainty in the intrinsic colour of the optical transient and the true extinction along the line of sight. The four integration time intervals, int1–int4, are measured in seconds from the Swift GRB trigger time at 01:42:18.7 UT on 2004 December 19. Notice the optical and γ -ray fluxes vary in concert so that the spectra never cross, and also that the highest-energy band seems to be a slightly better predictor of the behaviour of the optical emission.

the first measurement, any optical emission generated by internal shocks in GRB 990123 was outshone by bright optical emission from the external reverse shock. To generate the correlated optical and γ -ray variations measured throughout the full interval of γ -ray emission in GRB 041219a with internal (forward) shocks, reverse shock emission must be suppressed and/or delayed. The timing and strength of the reverse shock component depends strongly on the physical properties of the relativistic ejecta and the surrounding medium. In fact, the PAIRITEL near-infrared observations of GRB 041219a show the emergence of a weaker component after the end of the prompt γ -ray emission that can be interpreted as delayed reverse shock emission¹⁹.

With the addition of the new optical properties displayed by GRB 041219a to the set of known properties for optical emission from GRBs, we can construct a taxonomy of GRB optical emission with three classes: (1) prompt optical emission varying simultaneously with the prompt γ -rays; (2) early afterglow emission that may start during the prompt γ -ray emission, but persists for ten minutes or more after the prompt γ -rays have faded^{6,20–22}; and (3) late afterglow emission that can last for many hours to days^{23–25}. Within the context of the standard fireball model, it makes sense to attribute the prompt emission to internal shocks in the ultra-relativistic ejecta driven by the GRB engine², the early afterglow to a reverse

shock driven into the ejecta by interaction with the surrounding medium^{1–5,15}, and the late afterglow to forward external shocks driven into the surrounding medium generated by interaction with the ejecta^{26,27}. This theoretical framework, in turn, allows predictions about the timing, spectra and relative strength of the optical components that hinge on the properties of the inner engine, the ejecta and the surrounding medium. The ability of the Swift satellite to provide precise real-time positions and make panchromatic observations of GRBs²⁸, supplemented by a new generation of sensitive ground-based rapid response telescopes, therefore brings us into a new era in the study of the critical first few minutes during and after GRBs—one that will allow us to probe deeply the physics of these enigmatic explosions. □

Received 20 January; accepted 1 March 2005; doi:10.1038/nature03515.

1. Meszaros, P. & Rees, M. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–237 (1997).
2. Meszaros, P. & Rees, M. GRB 990123: reverse and internal shock flashes and late afterglow behaviour. *Mon. Not. R. Astron. Soc.* **306**, L39–L43 (1999).
3. Sari, R. & Piran, T. Predictions for the very early afterglow and the optical flash. *Astrophys. J.* **520**, 641–649 (1999).
4. Zhang, B., Kobayashi, S. & Meszaros, P. Gamma-ray burst early optical afterglows: Implications for the initial Lorentz factor and the central engine. *Astrophys. J.* **595**, 950–954 (2003).
5. Nakar, E. & Piran, T. Early afterglow emission from a reverse shock as a diagnostic tool for gamma-ray burst outflows. *Mon. Not. R. Astron. Soc.* **353**, 647–653 (2004).
6. Akerlof, C. *et al.* Observation of contemporaneous optical radiation from a gamma-ray burst. *Nature* **398**, 400–402 (1999).
7. Gotz, D., Mereghetti, S., Shaw, S., Beck, M. & Borkowski, J. GRB 041219 — A long GRB detected by INTEGRAL. *GRB Circ. Netw.*, 2866 (2004).
8. Barthelmy, S. *et al.* Swift-BAT detection of the bright long burst GRB 041219. *GRB Circ. Netw.*, 2874 (2004).
9. Paciesas, W. *et al.* The fourth BATSE gamma-ray burst catalog (revised). *Astrophys. J. Supp.* **122**, 465–495 (1999).
10. Vestrand, W. T. *et al.* The RAPTOR experiment: a system for monitoring the optical sky in real time. *Proc. SPIE* **4845**, 126–136 (2002).
11. Blake, C. & Bloom, J. S. GRB 041219: Infrared afterglow candidate. *GRB Circ. Netw.*, 2870 (2004).
12. Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust IR emission for use in estimation of reddening and CMBR foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
13. Hearty, F. *et al.* NIR observations of GRB 041219. *GRB Circ. Netw.*, 2916 (2004).
14. Henden, A. GRB 041219, BVR_I field calibration. *GRB Circ. Netw.*, 2871 (2004).
15. Panaitescu, A. & Kumar, P. Analysis of two scenarios for the early optical emission of the gamma-ray burst afterglow 990123 and 021211. *Mon. Not. R. Astron. Soc.* **353**, 511–522 (2004).
16. Kulkarni, S. R. *et al.* Discovery of a radio flare from GRB 990123. *Astrophys. J.* **522**, L97–L100 (1999).
17. Katz, J. I. Low-frequency spectra of gamma-ray bursts. *Astrophys. J.* **432**, L107–L109 (1994).
18. Liang, E. P., Crider, A., Bottcher, M. & Smith, I. A. GRB 990123: The case for saturated Comptonization. *Astrophys. J.* **519**, L21–L24 (1999).
19. Blake, C. H. *et al.* An infrared flash contemporaneous with the γ -rays from GRB 041219a. *Nature* doi:10.1038/nature03520 (this issue).
20. Vestrand, W. T. *et al.* RAPTOR: Closed-loop monitoring of the night sky and the earliest optical detection of GRB 021211. *Astron. Nachr.* **325**(6–8), 549–552 (2004).
21. Li, W., Filippenko, A., Chornock, R. & Jha, S. The early light curve of the optical afterglow of GRB 021211. *Astrophys. J.* **586**, L9–L11 (2003).
22. Fox, D. *et al.* Discovery of early optical emission from GRB 021211. *Astrophys. J.* **586**, L5–L8 (2003).
23. van Paradijs, J., Kovoveliotou, C. & Wijers, R. A. M. J. Gamma-ray burst afterglows. *Annu. Rev. Astron. Astrophys.* **38**, 379–425 (2000).
24. Stanek, K. Z. *et al.* Spectroscopic discovery of the supernova 2003dh associated with GRB 0303299. *Astrophys. J.* **591**, L17–L20 (2003).
25. Hjorth, J. *et al.* A very energetic supernova associated with the gamma-ray burst of 29 March 2003. *Nature* **423**, 847–850 (2003).
26. Waxman, E. Gamma-ray burst afterglow: supporting the cosmological fireball model, constraining parameters, and making predictions. *Astrophys. J.* **485**, L5–L8 (1997).
27. Sari, R., Piran, T. & Narayan, R. Spectra and light curves of gamma-ray burst afterglows. *Astrophys. J.* **497**, L17–L20 (1998).
28. Gehrels, N. *et al.* The Swift gamma-ray burst mission. *Astrophys. J.* **611**, 1005–1020 (2004).
29. Cenko, S. B. GRB 041219: optical afterglow detection. *GRB Circ. Netw.* 2886 (2004).
30. Fenimore, E. *et al.* Swift-BAT time history of GRB 041219. *GRB Circ. Netw.* 2906 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The RAPTOR project is supported by the Laboratory Directed Research and Development programme at Los Alamos National Laboratory.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.T.V. (vestrand@lanl.gov).

An infrared flash contemporaneous with the γ -rays of GRB 041219a

C. H. Blake¹, J. S. Bloom^{1,2}, D. L. Starr¹³, E. E. Falco³, M. Skrutskie⁹, E. E. Fenimore⁷, G. Duchêne¹², A. Szentgyorgyi⁵, S. Hornstein¹⁰, J. X. Prochaska⁴, C. McCabe¹¹, A. Ghez¹⁰, Q. Konopacky¹⁰, K. Stapelfeldt¹¹, K. Hurley⁵, R. Campbell⁶, M. Kassis⁶, F. Chaffee⁶, N. Gehrels⁸, S. Barthelmy⁸, J. R. Cummings⁸, D. Hullinger^{8,14}, H. A. Krimm^{8,15}, C. B. Markwardt^{8,14}, D. Palmer⁷, A. Parsons⁸, K. McLean⁷ & J. Tueller⁸

¹Harvard College Observatory, Cambridge, Massachusetts 02138, USA

²Astronomy Department, University of California at Berkeley, Berkeley, California 94720, USA

³Smithsonian Astrophysical Observatory, Cambridge, Massachusetts 02138, USA

⁴UCO/Lick Observatory, Santa Cruz, California 95064, USA

⁵Space Sciences Laboratory, University of California, Berkeley, California 94720, USA

⁶W.M. Keck Observatories, Kamuela, Hawaii 96743, USA

⁷Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

⁸NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

⁹University of Virginia, Department of Astronomy, Charlottesville, Virginia 22904, USA

¹⁰University of California, Los Angeles, Los Angeles, California 90095, USA

¹¹Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA

¹²Laboratoire d'Astrophysique, Observatoire de Grenoble, 38041 Grenoble Cedex 9, France

¹³Gemini Observatory, Hilo, Hawaii 96720, USA

¹⁴University of Maryland, College Park, Maryland 20742, USA

¹⁵Universities Space Research Association, Columbia, Maryland 21044, USA

The explosion that results in a cosmic γ -ray burst (GRB) is thought to produce emission from two physical processes: the central engine gives rise to the high-energy emission of the burst through internal shocking¹, and the subsequent interaction of the flow with the external environment produces long-wavelength afterglows^{2–4}. Although observations of afterglows⁵ continue to refine our understanding of GRB progenitors and relativistic shocks, γ -ray observations alone have not yielded a clear picture of the origin of the prompt emission⁶ nor details of the central engine. Only one concurrent visible-light transient has been found⁷ and it was associated with emission from an external shock. Here we report the discovery of infrared emission contemporaneous with a GRB, beginning 7.2 minutes after the onset of GRB 041219a (ref. 8). We acquired 21 images during the active phase of the burst, yielding early multi-colour observations. Our analysis of the initial infrared pulse suggests an origin consistent with internal shocks.

Prompt long-wavelength afterglow emission is predicted to arise when the reverse (external) shock encounters the ejecta of the explosion^{2,3}, or through γ -ray heating of the circumburst material⁴. Indeed, four GRBs^{7,9–12} have exhibited transient optical emission that could be associated with reverse shocks, but early-time optical transients have not been found for the vast majority of bursts (however, new larger-aperture robotic optical systems have met with increasing success). As even moderate levels of dust near the GRB or along the line-of-sight in the host galaxy could effectively suppress detectable optical emission¹³, contemporaneous observations at infrared (IR) wavelengths, where light suppression is relatively minimized, offer a natural means to uncover any prompt emission. This was one motivation for our construction of the Peters Automated Infrared Imaging Telescope (PAIRITEL; Fig. 1). At 1.3 m in diameter, it is one of the largest, completely autonomous telescope systems in the world and one of only a few capable of imaging at IR wavelengths (1.1–2.3 μ m; see ref. 14). PAIRITEL

acquires images with high temporal cadence (integration times of 7.8 s) in three colours simultaneously. The field of view is rather large, 8.5' $\times$ 8.5', for IR imaging, allowing for follow-up of GRB localizations of even modest precision.

GRB 041219a triggered the IBIS instrument on board the INTEGRAL¹⁵ satellite on 19 December 2004 at 01:42:54 UTC, which was reported at 01:44:05 UTC. The initial position of right ascension (RA) 00 h 24 min 26 s, declination (dec.) +62° 50' 06" was refined to 2' uncertainty at 01:47:49 UTC and a final offline location⁸ was reported at 03:31:58 UTC. The Burst Alert Telescope (BAT) on the Swift satellite¹⁶ triggered and located GRB 041219a on board at 01:42:18 UTC with a position that was within 4' of the IR source. If Swift had not still been in its commissioning phase with slewing disabled, the spacecraft could have slewed to the location within 70 s of the BAT trigger. The ground-based BAT location (RA 00 h 24 min 37.0 s, dec. +62° 50' 49.2") was within 48" of the IR source. As viewed by BAT, the burst duration (Δt) above background was 520 s and was very bright, with up to 6.5×10^4 counts s⁻¹ (unsaturated) between 15 and 350 keV and a fluence of 1.15×10^{-4} erg cm⁻². The time evolution of the count rate in four BAT channels covering 15 to 350 keV is reproduced in Fig. 2.

PAIRITEL began to slew on 19 December 2004, 01:48:20 UTC, and the first observations of the GRB field commenced 58 s later. Despite very poor observing conditions (sustained 40 m.p.h. winds, variable sky transmission, and 4" seeing), comparison of the first epoch of data revealed a new, variable source¹⁷ not visible in the 2MASS (Two Micron All Sky Survey) catalogue images of the field. When compared to the astrometric grid of 2MASS stars in the field, we find the absolute position of the IR transient (IRT) to be RA 00 h 24 min 27.68 s $\pm$ 0.124", dec. +62° 50' 33.501" $\pm$ 0.228", with its uncertainty dominated by the mapping to 2MASS catalogue stars. An optical flash was also detected by the RAPTOR experiment during the prompt γ -ray emission¹⁸.

PAIRITEL observations of the transient continued over the following three nights, until inclement weather in Arizona precluded additional observations. In total, 5,790 images were acquired by the system over these nights. In addition, we obtained deep J-band imaging on 20 and 21 December UTC using the NIRC-1¹⁹ instrument on the Keck I 10-m telescope on Mauna Kea, Hawaii. Consistent within the astrometric accuracy of the IRT from 19 December, we found a point-like source in J-band (Fig. 1). Over these two nights, that source was seen to fade by 1.0 mag, confirming its identification with the IRT. The Keck images revealed two sources within 2.5" of the transient position (S1; $J \approx 19.7$ mag, 2.5" north-north-east; S2: $J \approx 21.4$ mag, 1.5" east): both were unresolved apparent point sources. The source S1 is bright enough to contaminate the PAIRITEL J-band aperture photometry on 21 December, which accounts for the difference between our measurements and the fainter measurements²⁰ from Apache Point Observatory (APO) and Keck on the same date. When comparing PAIRITEL photometry with higher-resolution Keck and APO results for 21 December, the flux from S1+S2 appears to be a 51% contamination in H band, 58% in J band, and has negligible contribution in the K_s band.

The resulting light curves (see Supplementary Table 1) shown in Fig. 3 reveal a complex time history of the afterglow. The first six PAIRITEL exposures at $t + 7.2$ min after the trigger show a source that brightens, then fades very rapidly in all filters by about $t + 9$ min, and then rebrightens by $t + 20$ min. Using our data and data reported in the literature, we fitted the light curves as the sum of three smoothly connected rise and fall brightening events; the results of these fits are shown in Fig. 3. During the first few days, the source colours, though rather uncertain, appear consistent with a single value of the spectral slope of $\beta \approx 0.4$ (Fig. 4). Additionally, there is some evidence that the IRT was redder during the 'flash' event at $t + 7.2$ min.

How might the light curve be understood as emission from the reverse and forward shock? The electrons in the shock are assumed to be accelerated to a power-law spectrum with number density as a function of energy (E) proportional to E^{-p} . In a constant-density circumburst environment, a reverse shock is expected²¹ to rise rapidly ($\alpha = 3p - 3/2$, with flux density $f_\nu \propto t^\alpha$) and then, in the ‘thin shell’ case (see below), decline with $\alpha = -(27p + 7)/35 \approx -2$ after the emission peak, corresponding to the time ($t_\times$) that the reverse shock crosses the explosion ejecta. The IR emission from the forward shock²² should rise slowly as $\alpha = 0.5$, then decline as $\alpha = 3(1 - p)/4$. Associating peak 2 with the reverse shock and peak 3 with the forward shock, we find reasonable agreement, to within the measurement uncertainties, of the data with this model. In particular, the three implied values for p (2.5 ± 1.1 , 4.2 ± 4.1 , 2.6 ± 0.2 , for the reverse-rise, reverse-decline, and forward-decline, respectively) are all consistent with the usual range⁵ of $p = 2.2$ – 2.5 . A consequence of this interpretation is that, in the absence of effects due to collimation of the burst, radio afterglow emission should be dominated by a rising reverse shock, peaking at time $t = (10.2 \pm 1.8)(\nu/8.4)^{-35/54}$ d where ν is in GHz; this is thus far confirmed with reports of rising radio emission at least to day 2.9 (ref. 23).

With this interpretation, there are three puzzles. First, we would expect³ the source to have become bluer during the forward shock rise, which is not required (though is not excluded) by our data. The second puzzle concerns the timing of the reverse shock peak relative to the GRB duration. We expect^{24,25} $t_\times = 1,670$ s $> \Delta t$ only when Δt is less than the time when the shock begins to decelerate (commonly

deemed the ‘thin shell’ case). This deceleration time occurs when the shock has swept up from the circumburst environment a quantity $1/\Gamma_0$ times the entrained mass, where Γ_0 is the terminal Lorentz factor of the shock. A delayed reverse shock crossing time requires $\Gamma_0 < 67E_{52}^{1/8} n^{-1/8} (\Delta t/520)^{-3/8} [0.5(1+z)]^{3/8}$, where n is the circumburst particle density in units of baryons cm^{-3} , E_{52} is the energy in the shock in units of 10^{52} erg, Δt is in units of seconds and the redshift of the burst is z . This initial Lorentz factor constraint is uncomfortably smaller than the limits placed on previous values²⁵ of Γ_0 , which suggests that either the burst occurred at high redshift (which is excluded by the RAPTOR detection¹⁸), was exceptionally energetic, or occurred in a low-density environment. Alternatively, the observed variability could be due to inhomogeneous density structure in the circumburst environment or delayed energy injection into the blastwave.

The third and most intriguing puzzle is the physical origin of the IR flash at $t_1 = 462$ s that has marginal evidence for appearing more red than the afterglow at later times (here the subscript 1 refers to the first peak in the light curve). Regardless of the interpretation of events 2 and 3, the rapid rise and fall appear to preclude an association with a reverse shock: the fit of a power-law decay slope to the K_s data acquired less than 10 min from burst trigger yields $\alpha = -18 \pm 5$, whereas setting $\alpha = -2$, as expected of a reverse shock, yields an unacceptable fit (reduced $\chi^2 = 4.4$). The duration of the IR flash ($\delta t \approx 45$ s, taken as the full-width at half-maximum of the model fit to the data) is comparable to the widths of the largest timescale for substructure in γ -rays; indeed, this δt is remarkably similar to that of the optical pulse⁷ in GRB 990123. The

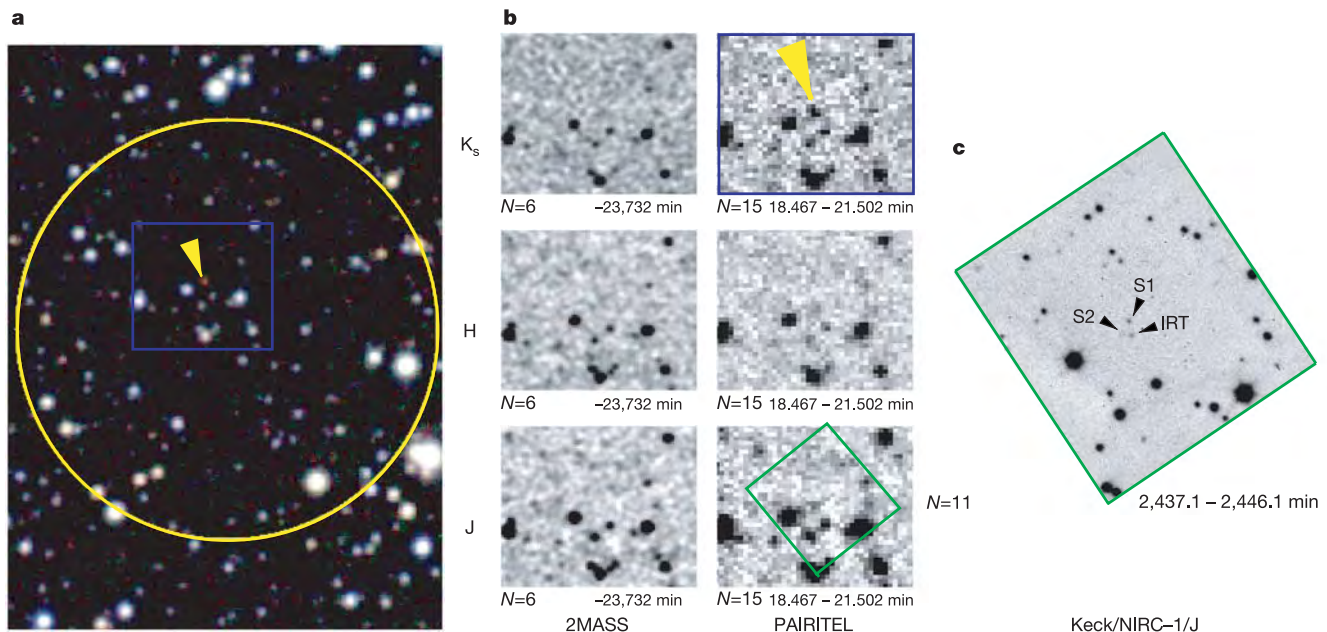


Figure 1 Images of the IR flash associated with GRB 041219a. The discovery was made with PAIRITEL, located atop Mt Hopkins, Arizona. Normally, the telescope’s observing plan is autonomously scheduled before nightfall by a routine that optimizes the preset priorities and scheduling constraints of all objects in the PAIRITEL database. When the new GRB alert was received, a series of observations were automatically inserted into the observing queue and the telescope began to slew to the target field. Left **a**, false-colour image of the IR flash (yellow arrow) of GRB 041219a inside the 2’ (90% confidence) INTEGRAL error region⁸ (yellow circle). At early times, the source is the reddest object in the field, indicative of very high extinction due to dust in the disk of our Galaxy. **b**, the Two Micron All Sky Survey (2MASS) images of the field from 2000 June 15 UTC compared with the three colour images showing the infrared transient (IRT) several minutes after the GRB

triggered. Time relative to the GRB trigger is given, as is the number (N) of individual images combined. The individual 2MASS images are 1.3-s integrations, so the combination of six 2MASS images is equal to a single PAIRITEL image. The 2MASS and PAIRITEL fields are centred on the IRT and are approximately 1’ on a side (shown as a blue box in **a**). A local background from a two-dimensional median was subtracted from all images to remove large-scale background variations. For all images, north is up and east is to the left. **c**, the Keck NIRC-1¹⁹ imaging at J-band on December 21 UTC, in the same region as the PAIRITEL images. NIRC-1 images were combined in the usual manner and the combined images had a seeing of less than 0.8” on both nights. The IRT, as well as two unresolved nearby sources (S1, S2), are labelled.

ratio of the width to the time after trigger $t \approx 462$ s, $\delta t/t = 0.10$, is similar to that seen in individual pulses in bright GRBs²⁶ but a factor of ~ 10 smaller than the $\delta t/t$ of the optical flash of GRB 990123.

We suggest, therefore, that the origin of the first peak is from the internal shock that itself produced the GRB. Such emission is possible if the synchrotron cooling frequency from the internal

shock emission is well below γ -ray frequencies³. Indeed, if the IR emission is due to internal shocks, then the observed flash may be due to a superposition of several unresolved shorter-timescale pulses. Future observations of IR flashes in the Swift era will no doubt test the ubiquity and nature of rapidly variable early-time emission as reported here.

Note added in proof: Some of our additional independent photometric analyses of the first event also reveal a rapid decline but show less evidence for a rapid rise than shown in Fig. 2. Since we based our conclusions on the nature of that event primarily on the later-time

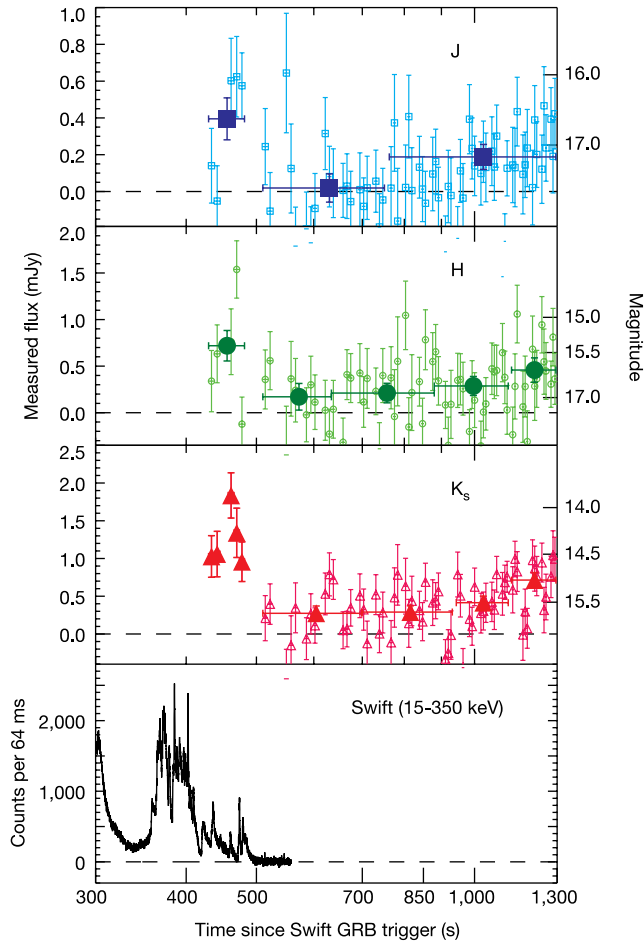


Figure 2 Evolution of the IR flash associated with GRB 041219a. We argue that the initial pulse is not traditional afterglow from an external shock but instead related to the internal shocks of the central engine. Shown (top three panels) are measurements derived from individual 7.8-s exposures as small, light-coloured points, and detections from stacks of images as heavy, dark-coloured points; error bars are 1σ , estimated from photon noise and the distribution of randomly placed apertures on the images. Also plotted (bottom panel) is the light curve from the 15–350-keV channels of the BAT instrument aboard Swift. Data reduction was as follows: the response to variable sky and bias in the detectors was estimated for each exposure in each band by median-combining all the exposures taken within a two to four minute window. A flat-field correction for fixed pixel-to-pixel variations in detector gain was made using images of the dawn sky. Subtracting the dark + sky response and normalizing by the flat-field produced a reduced image. Reduced images were measured either individually or as stacks of co-added images. As individual images undersample the seeing owing to atmospheric blurring, photometry was performed differentially in an aperture of fixed size. All of the images were aligned to a common reference image to an accuracy of approximately 0.1 pixel. Images were measured individually or in stacks created by summing individual images with weights determined by the signal-to-noise of each image. The measured flux at the position of the GRB was compared to the flux measured for a set of nearby comparison stars. Magnitudes of the comparison stars are known to a high accuracy (2%) from the 2MASS catalogue. PAIRITEL uses the same detectors and filters as 2MASS, so our differential photometry is expected to be free from systematic offsets.

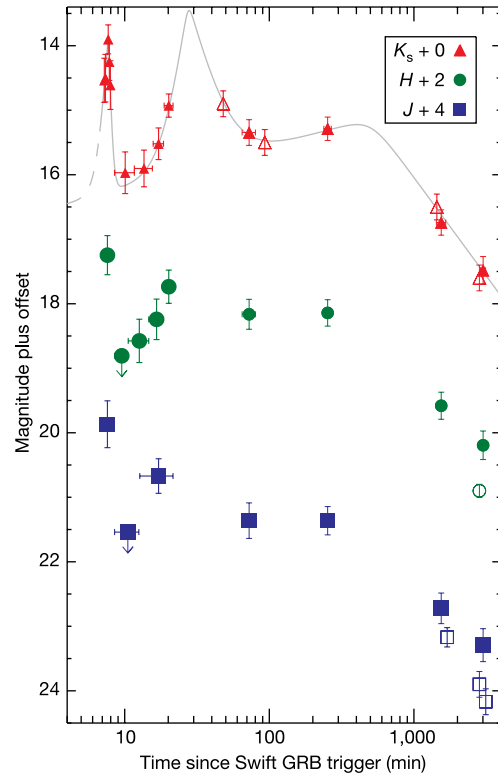


Figure 3 Long-timescale variability of the IR transient associated with GRB 041219a. The K_s -band light curve is reasonably fitted (grey curve) as a sum of three power-law rise and fall events with peaks at times since the burst $[t_1, t_2, t_3] = [7.73 \pm 0.14$ min, 27.5 ± 4.9 min, ≈ 500 min] and specific brightness $[f_\nu(t_1), f_\nu(t_2), f_\nu(t_3)] = [1.6 \pm 0.3$ mJy, 2.5 ± 2.8 mJy, 0.5 ± 0.1 mJy]. The rise and fall near the second peak is fitted by $\alpha_{2,r} = 6.1 \pm 2.9$ ($\alpha_{2,r} = -3.4 \pm 2.8$), with the specific brightness changing as $f_\nu \propto t^\alpha$. The final rise and fall parameters depend on the poorly constrained value of t_3 . Fixing $t_3 = 500$ min, we find $\alpha_{3,r} = 0.3 \pm 0.1$ and $\alpha_{3,f} = -1.2 \pm 0.1$. The first peak has both rapid rise and fall times (see text). The global χ^2 per degree of freedom (d.f.) is an acceptable 1.24. The reported 1σ parameter uncertainties do not reflect the covariance between the parameter fits. Additionally, all of these derived values depend on the choice of a smoothing parameter to connect the two power-laws (s in equation (2) of ref. 27; here, we chose $s = 5$). The J- and H-band light curves are consistent with the K_s -band shape, apart from the overall flux normalization. Photometry was performed as follows: large numbers of individual exposures were combined with sub-pixel sampling and produced a high signal-to-noise, high-resolution image for each epoch. The point-spread function in these images is well sampled, so the technique of difference image photometry was utilized. The uncertainty in the differential fluxes of the IRT between individual epochs is near photon-limited. The differential fluxes are converted to magnitudes from aperture photometry at the position of the IRT in the late-time (21 December) reference image. Other data, plotted as open points, include two J-band observations made with the Keck I telescope, JHK_s observations from Apache Point Observatory²⁰, and three K_s measurements (see ref. 28 and references therein) from Mt Palomar.

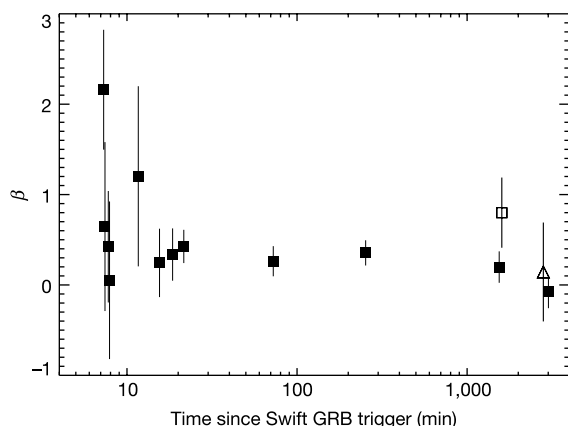


Figure 4 Colour evolution of the afterglow of GRB 041219a, corrected for extinction due to interstellar dust. The dust extinction²⁹ at the location of GRB 041219a is high: $A(V \text{ band}) = 5.9$ (and highly uncertain). Using a common dust extinction law ($R_V = 3.1$), the extinction in magnitudes is $A(J) = 1.6$, $A(H) = 1.04$, and $A(K_s) = 0.66$. We fitted (filled squares) for a spectral index β with our J, H and K_s measurements at each epoch using the flux relation $F_\nu(t) = F_0(t)(\lambda_c/\lambda_H)^\beta$, where λ_c is the central wavelength of the filter bandpasses ($c = [J, H, K_s]$). Including the z-band detection³⁰ ($\lambda_c \approx 9,100 \text{ \AA}$) that was nearly coeval with our IR observations results in a steeper spectral slope of 0.80 ± 0.39 (open square) than found with the IR data alone. Our fit to the Apache Point JHK data²⁰ is also shown with an open triangle. With fits that yielded $\chi^2/\text{d.f.}$ greater than unity, we scale the resulting 1σ errors by $\sqrt{\chi^2/\text{d.f.}}$. Though there is some evidence for secular trends, the resulting values for β are consistent with a single value of $\beta \approx 0.4$. We caution, however, that this common value of the spectral index is strongly dependent upon the assumed dust column. For different values of the extinction, $A(V) = 4.0$ and $A(V) = 7.0$, the fits to the spectral index are $\beta \approx 0.7$ and $\beta \approx 0.2$, respectively. Indeed, the standard forward-shock model²² tends to favour a larger value ($\beta \approx 0.55$) than found with the IR data alone. For the first peak (during the GRB), there is marginal evidence that the source is intrinsically redder than at later times.

imaging and the rapid fall behaviour, the interpretation is largely unchanged by various analyses. $\square$

Received 12 January; accepted 1 March 2005; doi:10.1038/nature03520.

1. Fenimore, E. E., Madras, C. D. & Nayakshin, S. Expanding relativistic shells and gamma-ray burst temporal structure. *Astrophys. J.* **473**, 998 (1996).
2. Mészáros, P. & Rees, M. J. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–237 (1997).
3. Sari, R. & Piran, T. Predictions for the very early afterglow and the optical flash. *Astrophys. J.* **520**, 641–649 (1999).
4. Beloborodov, A. M. Radiation front sweeping the ambient medium of gamma-ray bursts. *Astrophys. J.* **565**, 808–828 (2002).
5. van Paradijs, J., Kouveliotou, C. & Wijers, R. A. M. J. Gamma-ray burst afterglows. *Annu. Rev. Astron. Astrophys.* **38**, 379–425 (2000).
6. Baring, M. G. & Braby, M. L. A study of prompt emission mechanisms in gamma-ray bursts. *Astrophys. J.* **613**, 460–476 (2004).
7. Akerlof, C. et al. Observation of contemporaneous optical radiation from a gamma-ray burst. *Nature* **398**, 400–402 (1999).
8. Gotz, D., Mereghetti, S., Shaw, S., Beck, M. & Borkowski, J. GRB 041219—A long GRB detected by INTEGRAL. *GCN Circ.* 2866 (2004).
9. Fox, D. W. et al. Early optical emission from the γ -ray burst of 4 October 2002. *Nature* **422**, 284–286 (2003).
10. Fox, D. W. et al. Discovery of early optical emission from GRB 021211. *Astrophys. J.* **586**, L5–L8 (2003).
11. Li, W., Filippenko, A. V., Chornock, R. & Jha, S. The early light curve of the optical afterglow of GRB 021211. *Astrophys. J.* **586**, L9–L12 (2003).
12. Rykoff, E. S. et al. The early optical afterglow of GRB 030418 and progenitor mass loss. *Astrophys. J.* **601**, 1013–1018 (2004).
13. Klose, S. et al. The very faint K-band afterglow of GRB 020819 and the dust extinction hypothesis of the dark bursts. *Astrophys. J.* **592**, 1025–1034 (2003).
14. Zerbi, F. M. et al. *Proc. SPIE* **4841**, 737–748 (2003).
15. Winkler, C. et al. The INTEGRAL mission. *Astron. Astrophys.* **411**, L1–L6 (2003).
16. Gehrels, N. et al. The Swift gamma-ray burst mission. *Astrophys. J.* **611**, 1005–1020 (2004).
17. Blake, C. & Bloom, J. S. GRB 041219: Infrared afterglow candidate. *GCN Circ.* 2870 (2004).
18. Vestrand, W. T. et al. A link between prompt optical and prompt γ -ray emission in γ -ray bursts. *Nature* doi:10.1038/nature03515 (this issue).
19. Matthews, K., Soifer, B. T. & McLean, I. in *Infrared Astronomy with Arrays, the Next Generation* (ed. McLean, I.) 239 (Kluwer, Dordrecht, 1994).

20. Hearty, F. et al. NIR observations of GRB 041219. *GCN Circ.* 2916 (2004).
21. Kobayashi, S. Light curves of gamma-ray burst optical flashes. *Astrophys. J.* **545**, 807–812 (2000).
22. Sari, R., Piran, T. & Narayan, R. Spectra and light curves of gamma-ray burst afterglows. *Astrophys. J.* **497**, L17–L20 (1998).
23. van der Horst, A., Rol, E. & Strom, R. GRB 041219: Second epoch WSRT radio observations. *GCN Circ.* 2895 (2004).
24. Sari, R. Hydrodynamics of gamma-ray burst afterglow. *Astrophys. J.* **489**, L37–L40 (1997).
25. Zhang, B., Kobayashi, S. & Mészáros, P. Gamma-ray burst early optical afterglows: Implications for the initial Lorentz factor and the central engine. *Astrophys. J.* **595**, 950–954 (2003).
26. Ramirez-Ruiz, E. & Fenimore, E. E. Pulse width evolution in gamma-ray bursts: Evidence for internal shocks. *Astrophys. J.* **539**, 712–717 (2000).
27. Nakar, E. & Piran, T. Early afterglow emission from a reverse shock as a diagnostic tool for gamma-ray burst outflows. *Mon. Not. R. Astron. Soc.* **353**, 647–653 (2004).
28. Moon, D.-S., Cenko, B. & Adams, J. GRB041219: Continued NIR observations. *GCN Circ.* 2884 (2004).
29. Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
30. Cenko, S. B. GRB 041219: Optical afterglow detection. *GCN Circ.* 2885 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements J.S.B. was supported by a Junior Fellowship from the Harvard Society of Fellows. PAIRITEL was made possible by a grant from the Harvard Milton Fund. Additional funding from the Smithsonian Institution for the PAIRITEL project is acknowledged. We thank the entire Mt Hopkins Ridge staff for support of PAIRITEL, especially W. Peters, R. Hutchins and T. Groner. J. Huchra is thanked for stewardship of the telescope under adverse conditions in the week leading up to GRB 041219a. This publication makes use of data products from the Two Micron All Sky Survey (2MASS), which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S.B. (jbloom@astro.berkeley.edu).

Albedo of the south pole on Mars determined by topographic forcing of atmosphere dynamics

Anthony Colaprete¹, Jeffrey R. Barnes², Robert M. Haberle¹,
Jeffery L. Hollingsworth³, Hugh H. Kieffer⁴ & Timothy N. Titus⁴

¹NASA Ames Research Center, Space Science Division, Moffett Field, Mountain View, California 94035, USA

²College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA

³San José State University Foundation, NASA Ames Research Center, Moffett Field, Mountain View, California 94035, USA

⁴US Geological Survey, Astrogeology Team, Flagstaff, Arizona 86001, USA

The nature of the martian south polar cap has remained enigmatic since the first spacecraft observations^{1–6}. In particular, the presence of a perennial carbon dioxide ice cap, the formation of a vast area of black ‘slab ice’ known as the Cryptic region and the asymmetric springtime retreat of the cap have eluded explanation. Here we present observations and climate modelling that indicate the south pole of Mars is characterized by two distinct regional climates that are the result of dynamical forcing by the largest southern impact basins, Argyre and Hellas. The style of surface frost deposition is controlled by these regional climates. In the cold and stormy conditions that exist poleward of 60° S and extend 180° in longitude west from the Mountains of Mitchell (~30° W), surface frost accumulation is dominated by precipitation. In the opposite hemisphere, the polar atmosphere is relatively warm and clear and frost accumulation is dominated by direct vapour deposition. It is the differences in these deposition styles that determine the cap albedo.

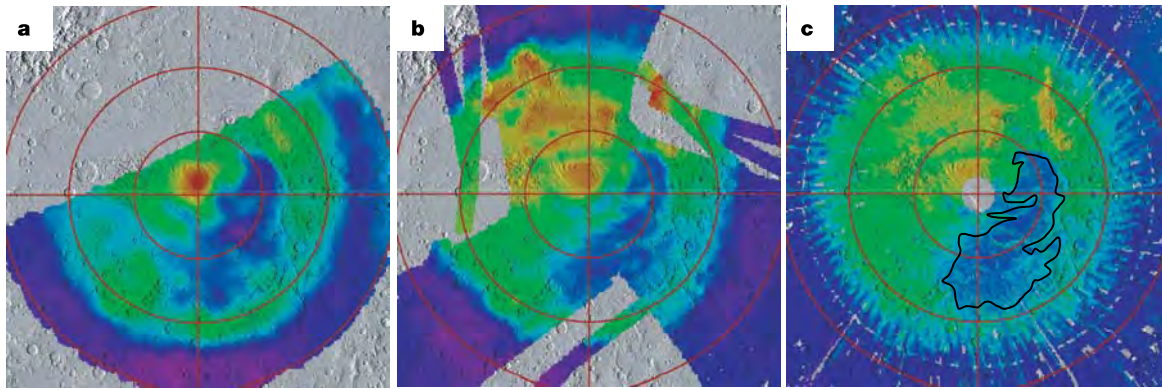


Figure 1 South pole springtime albedo maps from the Viking and Mars Global Surveyor spacecraft. Three albedo maps of the martian south springtime pole overlain on MOLA topography. In each panel, high albedos are indicated by warm colours (red shading) and low albedos are indicated by cool colours (blue shading). The red lines indicate every 10° of latitude and 90° of longitude. At the time of these observations the seasonal polar cap, composed primarily of CO₂ ice, extends from the pole to approximately 65° S. **a**, Albedo

measured by the Infrared Thermal Mapper on board the Viking spacecraft in 1976.

b, c, Albedo derived from observations made by the Thermal Emission Spectrometer on board the Mars Global Surveyor (MGS) spacecraft in 1999 (**b**) and 2001 (**c**). The black curve in **c** indicates the approximate boundary of the Cryptic region. Comparing **a** with **b** and **c**, there appear to be only minor differences in albedo between the Viking and MGS eras, with the largest differences occurring in parts of the Cryptic region.

The observed springtime (solar longitude, $L_s \approx 200$) surface albedo in the martian southern polar region is shown in Fig. 1. The Hellas impact basin is centred at -25° S and 65° W. In general, the hemisphere west of Hellas ('the western hemisphere') is marked by relatively high values of surface albedo. In contrast, the hemisphere east of Hellas ('the eastern hemisphere') contains extensive regions of very low surface albedo. One of the brightest features within the western hemisphere is the south pole residual cap (SPRC). The dark region that dominates the eastern hemisphere is the Cryptic region⁵.

The nature of the SPRC has been the source of considerable debate since its identification as CO₂ ice by the Viking spacecraft. Two fundamental questions still exist regarding the SPRC's formation, location and stability. First, why is the SPRC offset from the geographic pole? There are no local topographic features or surface properties that can account for the offset in the SPRC. Second, does the SPRC represent a large or a small reservoir of CO₂? If the former, then it could possibly buffer the surface pressure¹. If the latter, then the SPRC may not survive every year.

Just as mysterious as the SPRC is the region of 'black' CO₂ ice that dominates the eastern hemisphere (Fig. 1). Known as the Cryptic region⁵, this area, first observed by Viking and later by Mars Global Surveyor (MGS), becomes visible during the recession of the

seasonal cap. The Cryptic region is characterized by its very low visible albedo (<0.2) and high 25- μ m relative emissivity (>0.95) and is believed to be composed of very clean CO₂ 'slab' ice. It is the Cryptic region that first sublimates in spring, leading to the observed asymmetric retreat of the southern seasonal cap.

The difference in albedo between the SPRC and the Cryptic region suggests two styles of frost emplacement, and yet, to date, no obvious correlation between topography and surface properties that would favour a SPRC or Cryptic region have been identified⁵. However, independent observations from MGS indicate that the atmospheric circulation within the polar night produces two distinct climates that encompass the SPRC and the Cryptic region. These observations include cloud echoes from the Mars Orbiter Laser Altimeter (MOLA), ice grain sizes from the Thermal Emission Spectrometer (TES) and temperatures from TES and MGS Radio Science (RS).

This work, using simulations with the NASA Ames Mars General Circulation Model (MGCM), provides a consistent explanation for the various observations and offers a simple interpretation—the topography associated with the Tharsis rise and the Hellas and Argire impact basins modulates the south polar circulation, producing two distinct climates. Ultimately, the two climates generated by the circulation result in very different surface properties for the CO₂

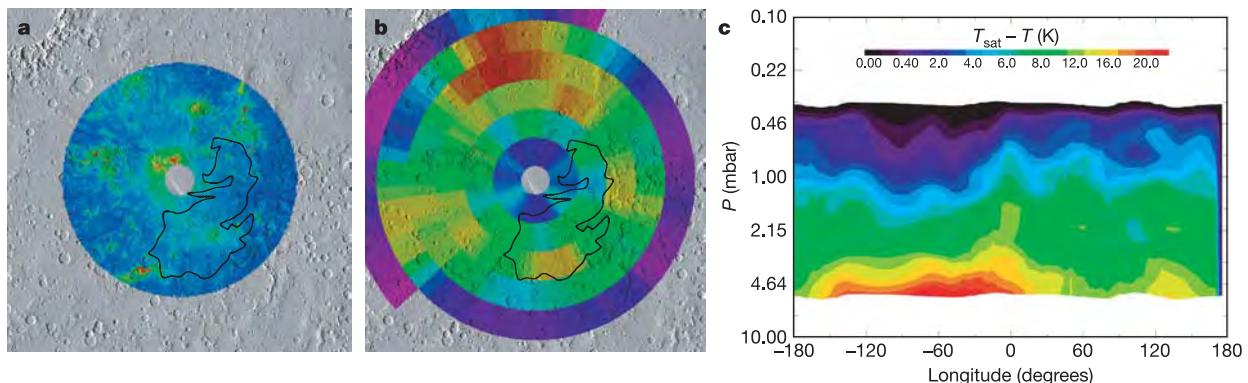


Figure 2 Observations from MGS showing south pole climate asymmetry in grain size and $\delta T_{\text{sat}} = T_{\text{sat}} - T$. **a**, Average grain size index from $L_s = 87$ –110. High values of the grain index (smaller grains) are indicated by warm colours (red shading), and low values of the grain index (larger grains) are indicated by cool colours (blue shading). **b**, Column-

integrated δT_{sat} derived from TES nadir atmospheric temperatures for the period $L_s = 90$ –110. The black contour shows the approximate boundary of the Cryptic region (Fig. 1). **c**, Longitude cross-section at 75° S showing δT_{sat} derived from TES as a function of pressure.

ice cap. Within the hemisphere that includes the SPRC, atmospheric precipitation dominates direct surface deposition, resulting in high albedo frosts; in the Cryptic region, cap ice is accumulated by direct deposition, forming dark CO₂ slab ice.

The albedo and emissivity of the polar cap has been associated with the size of the ice grains that make up the surface ice deposits^{5,7,8}. Small ice grains result in brighter surfaces, while larger grains result in darker surfaces. A measure of the ice grain size can be derived from differences in the 18- μ m and the 25- μ m brightness temperatures observed by the TES. The 18- μ m brightness temperature is a good measure of the true kinetic temperature, while the 25- μ m band is very sensitive to grain size. The 'grain size index' (Fig. 2a) is the difference between the two bands ($T_{18} - T_{25}$). Regions with high values of the grain size index include the SPRC and the Mountains of Mitchel. It has been suggested that high grain

size indexes are the result of small ice grains in the atmosphere or freshly precipitated to the surface⁸. MOLA detected cloud echoes within the polar night that have been attributed to the presence of CO₂ clouds⁹⁻¹¹. The greatest rates of occurrence are in the western hemisphere. In the eastern hemisphere there are fewer cloud echoes, with a minimum poleward of Hellas basin¹².

Also correlating with surface albedos are temperatures derived from TES and RS. Frequently the atmosphere within the polar night chills to below the condensation temperature of CO₂ ice and becomes supersaturated. The greatest and most persistent supersaturation, $\delta T_{\text{sat}} = T_{\text{sat}} - T$, occurs in the western hemisphere and in the longitude corridor of the SPRC (Fig. 2b). Atmospheric temperatures are significantly colder in the western hemisphere compared with the eastern hemisphere. The greatest values of δT_{sat} correlate well with surface brightness and grain size.

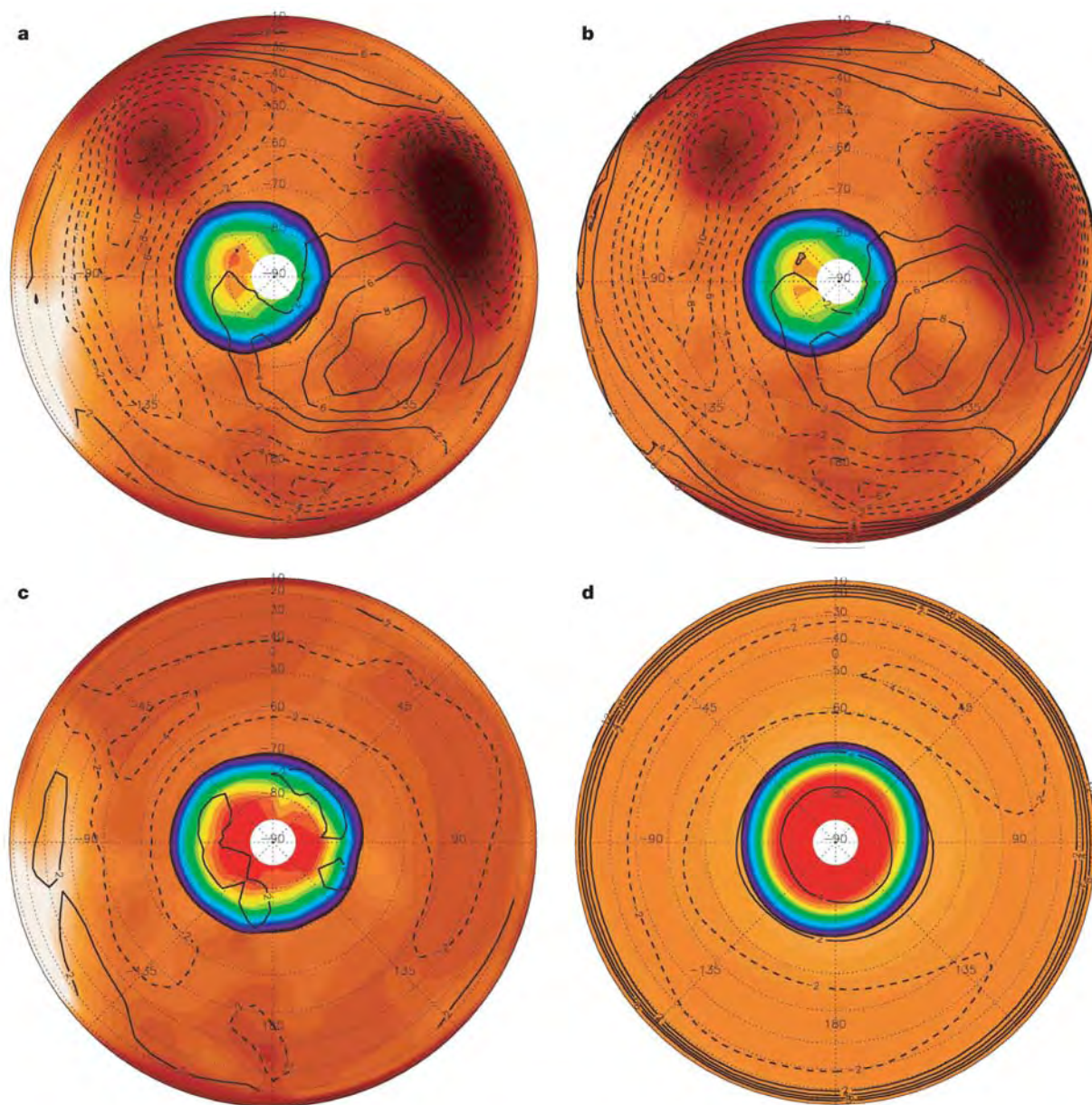


Figure 3 GCM results showing the dependence of the climate asymmetry on mid-latitude topography. All panels show NASA Ames MGCM simulation results for column-integrated δT_{sat} (colour shading with values corresponding to the colour bar shown in Fig. 2) at $L_s = 90$. Black contours indicate temperature deviations from the time-averaged (20 solar days) zonal mean at the 3.5-mbar pressure level, with solid contours indicating

temperatures higher than the zonal mean (high-pressure regions) and dashed contours indicating temperatures lower than the zonal mean (low-pressure regions). Model topography is indicated with the dark red–white background shading: **a**, Current true topography; **b**, Tharsis removed; **c**, Hellas and Argyre removed; **d**, Flat topography.

It is the atmospheric temperatures that provide insight into what is controlling these two climate regimes. The δT_{sat} for a longitude cut at constant latitude of 75° S is shown in Fig. 2c. Evident in the temperature deviations (Fig. 2b, c) is a low-level planetary wave with a very strong stationary zonal wavenumber-one component. Hinson *et al.*¹³ had previously identified this stationary wave in RS and TES observations. The wavenumber-one pattern establishes a low-pressure zone over much of the western hemisphere and a high-pressure zone over much of the eastern hemisphere. This wavenumber-one pattern dominates most of the southern winter, and extends into southern spring ($L_s \approx 230$). The strong wavenumber-one pattern results in a polar vortex that is offset from the geographic pole and sits over the SPRC.

Furthermore, the amount of variance in atmospheric temperatures is also greatest in the western hemisphere. A distinct 'storm track' exists, starting west of Hellas and extending west past Argyre¹⁴. Within this storm track temperatures can vary by as much as 15–20 K per solar day. The rapid decreases in temperature drive the formation of thick clouds and increased precipitation.

To further understand the southern winter dynamics and their association with the SPRC and Cryptic region, several simulations were conducted with the NASA MGCM¹⁵. The version of the MGCM used here includes a sophisticated CO₂ cloud scheme that predicts cloud formation and precipitation¹⁶. Simulations were conducted for a variety of topographic permutations in which the southern extent of Tharsis, Hellas and Argyre were included or removed (Fig. 3).

In the simulation with true (that is, current MOLA observed) topography, temperatures predicted by the model closely match those observed and, in particular, the coldest temperatures are observed in the western hemisphere and encompass the SPRC. Predicted CO₂ cloud cover, precipitation and δT_{sat} agree well with the observations described so far (compare Figs 3a and 2b). During the formation of the seasonal cap, cloud cover and precipitation are greatest in a longitude corridor that aligns with the SPRC and are almost completely absent over the Cryptic region. A strong wavenumber-one system dominates from late autumn to mid-winter and establishes a pattern of low and high surface pressures that are consistent with the observations.

When the topography is altered, the regional circulation and, hence, atmospheric temperatures are changed. Figure 3b shows the effect of removing the southernmost extent of the Tharsis rise. Without the southern extent of Tharsis, the hemispherical asymmetry in δT_{sat} , while still evident, is rotated west by about 45° and has somewhat smaller values compared with the case with true topography. If the Hellas and Argyre impact basins are removed (Fig. 3c), the wavenumber-one circulation pattern no longer dominates and cold temperatures occur with much greater symmetry about the pole. A very symmetrical pattern of circulation and cold temperatures results if all topography is removed (Fig. 3d). The simulations indicate that the south polar circulation is controlled by a quasi-stationary planetary (that is, Rossby) wave¹⁷ that produces longitudinal variation in the time-mean (for example, seasonal-mean) circulation and climate. The longitudinal asymmetries of the order of 20–30 K seen in the thermal environment between the western and eastern hemispheres in southern high latitudes are a result of this planetary wave that is excited primarily by the planet's middle-latitude surface topography distribution^{18,19}.

The two climates established by the topographically forced circulation produce differences in the nature of the surface ice deposits. From the MGCM simulations, the relative amounts of CO₂ ice that is deposited by precipitation and by direct deposition can be compared. The deposition ratio is defined as the amount of atmospheric precipitation ($\text{kg m}^{-2} \text{s}^{-1}$) divided by the amount of direct surface deposition of CO₂ ice. The deposition ratio is area-weighted so that longitude corridors of differing extent can be compared. On average, across the entire polar cap, the deposition

ratio is approximately 0.1. However, in some regions the deposition ratio can be as high as 0.9. The greatest deposition ratios occur in the western hemisphere. In the 'true topography' case of the MGCM, the deposition ratio between the longitudes of –135° W and 45° W is 1.56 times as large as the deposition ratio in the opposite hemisphere. If the western edge of this corridor is extended to –180° W, so that the western corridor now extends from –180° W to 45° W, and the eastern hemisphere corridor is confined to just the longitude extent of the Cryptic region, the deposition ratio is 22.4 times as large.

Jakosky and Haberle²⁰ suggested that the SPRC may be unstable from year to year, and that small perturbations in climate may result in the SPRC disappearing some years and reappearing in others. In spring, the recession of the seasonal cap is largely controlled by the surface brightness. Although albedos in local regions on the seasonal cap can change rapidly, the persistent asymmetric circulation creates conditions in which relatively larger grain sizes are stable in the Cryptic region through the sublimation season. Any changes to the circulation can result in changes to the surface brightness and poleward heat flow, and thus modify the recession of the cap. The current SPRC albedo is very close to the minimum required for year-round stability²⁰. If these albedos are only slightly modified by changes in the location or quantity of precipitation, then the SPRC will undergo significant changes and may partially or completely disappear. The observed changes in polar ice features reported by Malin *et al.*²¹ and Byrne and Ingersoll^{22,23} are consistent with the notion of a SPRC that is only quasi-stable and may undergo significant changes from one year to the next.

The albedo of the south polar ice cap is primarily controlled by the frost grain size. Frost grain size is determined by the style of frost deposition. Smaller grains form when there is atmospheric precipitation to the surface and larger grains form when frost is directly deposited to the surface. Locations that favour one style of frost emplacement over the other are primarily determined by the regional circulation and resulting thermal state. Thus topography, through its influence on the circulation, is ultimately responsible for the nature and location of the SPRC and Cryptic regions on Mars. □

Methods

The NASA Ames MGCM simulations presented here were run at a resolution of 5° latitude × 6° longitude. A constant atmospheric dust optical depth of 0.3 was assumed in all runs. Included in these simulations is a sophisticated carbon dioxide cloud scheme that models the microphysical cloud processes of nucleation, condensation/sublimation, and sedimentation¹⁶. There are several significant improvements in this MGCM CO₂ cloud model over previous models. Most relevant to the work presented here is the treatment of the nucleation and growth processes. The formation of new cloud particles occurs through the heterogeneous nucleation on dust or water-ice particles. A condensational growth rate is calculated at each time step (that is, it is not assumed to be instantaneous) and cloud particles are treated as tracers with their sedimentation rate calculated across the distribution of particle sizes. To accurately treat the growth of both cloud particles and direct deposition of surface ice, CO₂ vapour is treated as a tracer, independent of the total pressure. These features of the cloud model allow for supersaturation, as seen in the observations, and provide a much better representation of actual polar processes including precipitation and direct surface deposition.

Received 25 October 2004; accepted 10 March 2005; doi:10.1038/nature03561.

- Leighton, R. L. & Murray, B. C. Behaviour of carbon dioxide and other volatiles on Mars. *Science* **153**, 136–144 (1966).
- Paige, D. A. & Ingersoll, A. P. Annual heat balance of Martian polar caps Viking observations. *Science* **228**, 1160–1168 (1985).
- Lindner, B. L. The hemispherical asymmetry in the Martian polar caps. *J. Geophys. Res.* **98**, 3339–3344 (1993).
- James, P. B. *et al.* Seasonal recession of Martian South Polar Cap; 1992 HST observations. *Icarus* **123**, 87–100 (1996).
- Kieffer, H. H., Titus, T. N., Mullins, K. F. & Christensen, P. R. Mars south polar spring and summer behaviour observed by TES: Seasonal cap evolution controlled by frost grain size. *J. Geophys. Res.* **105**, 9653–9700 (2000).
- Thomas, P. C. *et al.* North–south geological differences between the residual polar caps on Mars. *Nature* **404**, 161–164 (2000).
- Forget, F., Houdin, F. & Talagrand, O. CO₂ snowfall on Mars: Simulation with a general circulation model. *Icarus* **131**, 302–316 (1998).
- Titus, T. N., Kieffer, H. H., Mullins, K. F. & Christensen, P. R. TES remapping data: Slab ice and snow

- flurries in the Martian north polar night. *J. Geophys. Res.* **106**, 23181–23196 (2001).
9. Pettengill, G. H. & Ford, P. G. Winter clouds over the north Martian polar cap. *Geophys. Res. Lett.* **27**, 609–613 (2001).
 10. Ivanov, A. B. & Muhleman, D. O. Cloud reflection observations: results from the Mars Orbiter Laser Altimeter. *Icarus* **154**, 190–206 (2001).
 11. Colaprete, A. & Toon, O. B. Carbon dioxide snow storms during the polar night on Mars. *J. Geophys. Res.* **107**, doi:10.1029/2001JE001758 (2002).
 12. Colaprete, A., Haberle, R. M. & Toon, O. B. Formation of convective carbon dioxide clouds near the south pole of Mars. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002053 (2003).
 13. Hinson, D. P., Wilson, R. J., Smith, M. D. & Conrath, B. J. Stationary planetary waves in the atmosphere of Mars during southern winter. *J. Geophys. Res.* **108**, doi:10.1029/2002JE001949 (2003).
 14. Hinson, D. P. & Wilson, R. J. Transient eddies in the Southern Hemisphere of Mars. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014103 (2002).
 15. Haberle, R. M. *et al.* General circulation model simulations of the Mars Pathfinder atmospheric structure investigation/meteorology data. *J. Geophys. Res.* **104**, 8957–8974 (1999).
 16. Colaprete, A. & Toon, O. B. Carbon dioxide clouds in an early dense Martian atmosphere. *J. Geophys. Res.* **108**, doi:10.1029/2002JE001967 (2003).
 17. Holton, J. R. *An Introduction to Dynamic Meteorology* 3rd edn (Academic, London, 1992).
 18. Hollingsworth, J. L. & Barnes, J. R. Forced, stationary planetary waves in Mars' winter atmosphere. *J. Atmos. Sci.* **53**, 428–448 (1996).
 19. Banfield, D., Conrath, B. J., Smith, M. D., Christensen, P. R. & Wilson, R. J. Forced waves in the Martian atmosphere from MGS TES nadir data. *Icarus* **161**, 319–345 (2003).
 20. Jakosky, B. M. & Haberle, R. M. Year-to-year instability of the Mars south polar cap. *J. Geophys. Res.* **95**, 1359–1365 (1990).
 21. Malin, M. C., Caplinger, M. A. & Davis, S. D. Observational evidence for an active surface reservoir of solid carbon dioxide on Mars. *Science* **294**, 2146–2148 (2001).
 22. Byrne, S. & Ingersoll, A. P. A sublimation model for Martian south polar ice features. *Science* **299**, 1051–1053 (2003).
 23. Byrne, S. & Ingersoll, A. P. Martian climate events on timescales of centuries: Evidence from feature morphology in the residual south polar ice cap. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017597 (2003).

Acknowledgements We acknowledge discussions with F. Montmessin and F. Forget. This work was supported under NASA's Planetary Atmospheres Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.C. (Anthony.Colaprete-1@nasa.gov).

Friction enhances elasticity in granular solids

C. Goldenberg^{1*} & I. Goldhirsch²

¹School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, ²Department of Fluid Mechanics and Heat Transfer, Faculty of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel

* Present address: Physique et Mécanique des Milieux Hétérogènes, ESPCI, 10 rue Vauquelin, 75231 Paris Cedex 5, France

For years, engineers have used elastic and plastic models to describe the properties of granular solids, such as sand piles and grains in silos^{1–3}. However, there are theoretical^{4–6} and experimental^{7–14} results that challenge this approach. Specifically, it has been claimed^{4–6} that stress in granular solids propagates in a manner described by wave-like (hyperbolic) equations, rather than the elliptic equations of static elasticity. Here we report numerical simulations of the response of a two-dimensional granular slab to an external load, revealing that both approaches are valid—albeit on different length scales. For small systems that can be considered mesoscopic on the scale of the grains, a hyperbolic-like, strongly anisotropic response is expected. However, in large systems (those typically considered by engineers), the response is closer to that predicted by traditional isotropic elasticity models. Static friction, often ignored in simple models, plays a key role: it increases the elastic range and renders the response more isotropic, even beyond this range.

Collections of macroscopic particles, known as granular

materials, are commonplace in the energy, pharmaceutical, chemical and food industries, as well as in the environment. Under the influence of gravity, grains organize in static assemblies. Understanding the properties of these granular solids is of practical and scientific importance. Traditionally, engineers have used elastic theory^{1,2} to describe granular bulk properties at low loads, and elasto-plastic models^{1–3} to describe yield and quasi-static flow. Surprisingly, in some experiments the pressure on the floor under sand piles was found to possess a local minimum under the apex^{13–14}, whereas isotropic elastic theory (and common sense) would imply that the pressure is maximal there (we note that the presence of this minimum depends on the method of preparation of the pile¹⁵). Other experiments^{7–9} revealed the existence of anisotropic chains of particle contacts ('force chains') along which the contact forces were stronger than average. These and other experiments^{10–12} were interpreted as evidence against elastic theories of granular matter. Models^{4–6} in which the forces propagate, like waves, along force chains, were proposed and some of their predictions were found to be in good agreement with experiment. The basic equations in these theories are hyperbolic, in contrast with the elliptic nature of static elasticity, implying a possible need to revisit important parts of engineering and environmental science. Here we show that this is not required, mostly because the new approaches (which often ignore interparticle friction⁶) apply mainly to mesoscopic-sized systems. An understanding of the crossover from the small- to the large-scale granular physics is essential for a better understanding of granular matter in general.

Consider a vertical force applied to the top of a rectangular layer ('slab') of a granular material, positioned on a horizontal surface. In a hyperbolic model^{4–6}, the forces propagate along a cone, or two lines in two dimensions (2D), so that the maximal pressure on the floor is not below the point of application of the force, but rather on a ring whose diameter increases with the depth of the layer (or two points, in 2D). The pressure distribution at the floor therefore exhibits two peaks (in 2D). In contrast, isotropic elasticity predicts a single peak, whose width is proportional to the depth of the layer^{16,17}, as observed in some experiments^{16,18}. Other experiments appear to be compatible with a hyperbolic description (for example, for lattice configurations^{8–10}). Experiments therefore seem to be providing mixed messages on the correct description of granular statics.

As two-dimensional simulations are easier to perform and visualize, we set out to study the microscopic basis of the above-mentioned findings in 2D. We studied two-dimensional slabs, composed of monodisperse or polydisperse disks, using a standard grain interaction model¹⁹: the normal and tangential forces, denoted by F_n and F_t respectively, are determined by spring-dashpot models (with spring constants k^n and k^t , respectively). A transition to sliding occurs when the Coulomb friction law is saturated: $F_t = \mu F_n$, where μ is the coefficient of static friction. Our numerical experiments proceeded as follows. First, an initial configuration of non-touching grains is prepared in an enclosure composed of rigid side walls and floor. The system is then relaxed under gravity to form a stationary slab (for details, see Fig. 1 legend). A downward vertical force (F_{ext}) is slowly applied to the top of the system until it reaches the desired value; the system is then further relaxed to a numerically static state.

Consider first a monodisperse system, with the disks arranged on a triangular lattice. The force profile on the floor for different values of F_{ext} is shown in Fig. 1a–c, for systems with different coefficients of friction. A crossover from a single-peaked to a two-peaked response occurs as the applied force is increased (except for $\mu = 1$), the crossover value of the force increasing with μ . The dependence of the value of the force on the floor below the point of application of F_{ext} ($x = 0$) on F_{ext} is shown in Fig. 1d. An elastic-like response (linear in F_{ext}) is observed for small F_{ext} . The extent of the range in which the response is linear in F_{ext} is significantly larger in frictional systems. As we have verified that superposition holds as

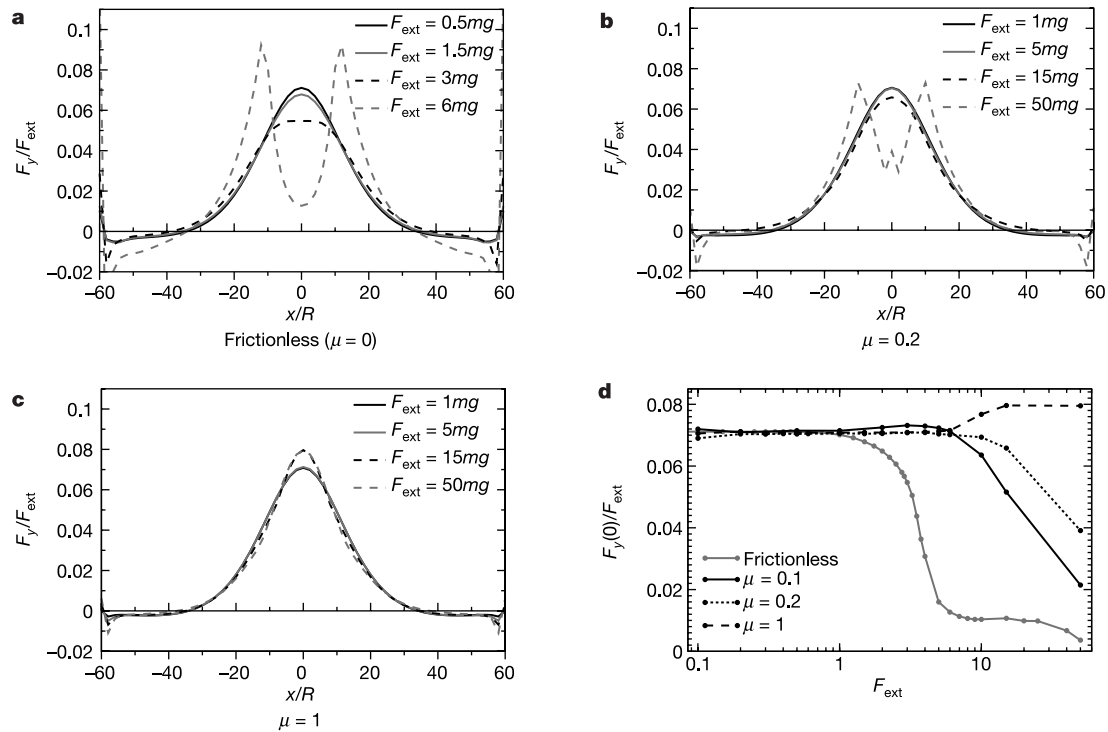


Figure 1 Force distribution and linearity below a lattice of disks. The vertical force profile on the floor of a triangular lattice of disks (60 particles wide and 15 deep), in response to different values of a vertical applied force, F_{ext} , at the top at $x = 0$; the effect of gravity is subtracted. These are results of numerical simulations with the following parameters (also used for polydisperse systems): the normal spring constant for particle–particle interactions, $k^n = 3,000\bar{m}g/\bar{R}$, where $\bar{R}$ and $\bar{m}$ are the mean particle radius and mass and g is the gravitational acceleration, and for particle–wall interactions $k_{\text{wall}}^n = 2k^n$. The ratio of tangential to normal stiffness (k^t/k^n), for $\mu \neq 0$, is 0.8. The time step used is

$5 \times 10^{-4} \sqrt{\bar{R}/g}$. The force is increased from 0 to F_{ext} in a time $\sqrt{\bar{R}/g}$; typically the relaxation to a numerically static state (in which we verified that the force response converges to better than 1% accuracy) requires a time between $\sqrt{\bar{R}/g}$ and $30\sqrt{\bar{R}/g}$, leading to a kinetic energy per particle of $10^{-13}\bar{m}g\bar{R}$ (and $10^{-9}\bar{m}g\bar{R}$ in the polydisperse case). Note that because the viscous-like damping used here is an artificial device designed to achieve a static state, it is questionable whether this time corresponds to a realistic physical time. **a**, $\mu = 0$ (frictionless). **b**, $\mu = 0.2$. **c**, $\mu = 1$. **d**, The vertical force on the floor at $x = 0$ with increasing F_{ext} , for different values of μ .

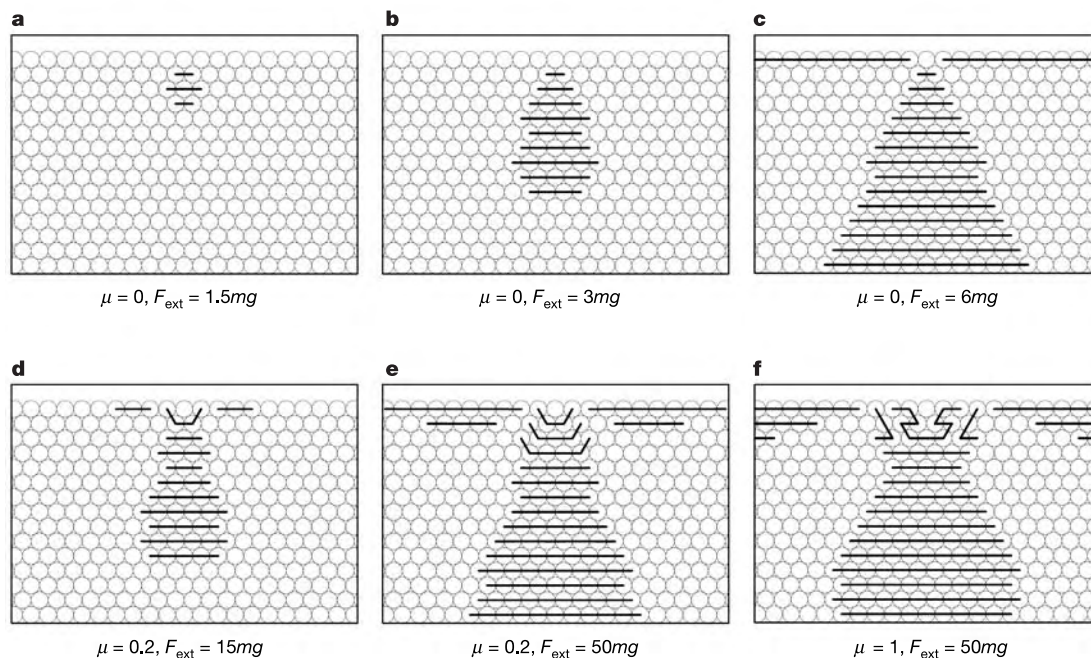


Figure 2 The changes (the teardrop) in the contact network induced by an applied force, for different values of F_{ext} and μ . Particles are indicated by circles, and lines connecting particle centres indicate open contacts (compared to the case when gravity alone is

present and no external force is applied; the loss of contact that follows slipping is not separately marked, for simplicity). Only the central third of the system width is shown.

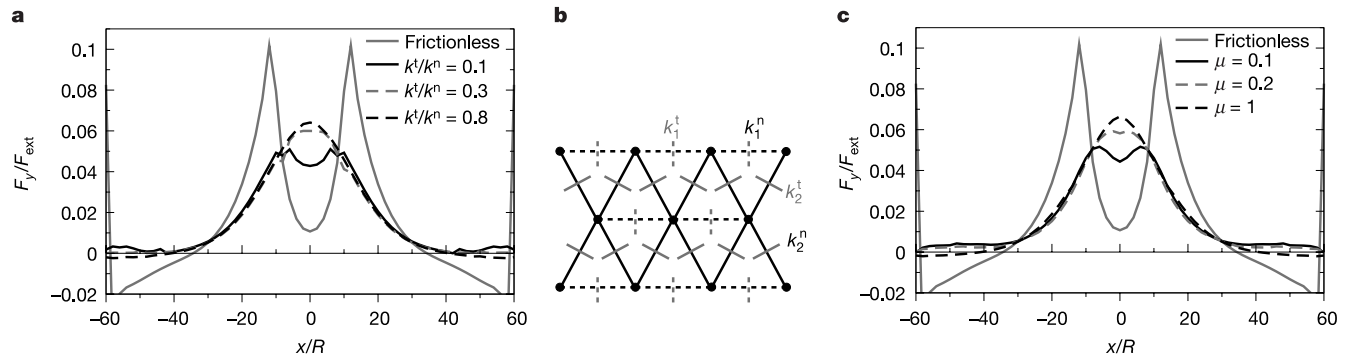


Figure 3 Effects of friction. **a**, The response of an ordered triangular lattice of disks with $F_{\text{ext}} = 15mg$, for different values of k^t/k^n and $\mu = 10$ (practically equivalent to $\mu = \infty$). **b**, A model spring network with normal and tangential springs. Nearest neighbours connected by normal springs with stiffness k_1^n , k_2^n for horizontal and oblique springs,

respectively, and by tangential springs with stiffness k_1^t , k_2^t , respectively. The rest length of the normal springs is taken to equal the lattice constant, and that of the tangential springs is zero. **c**, Same as **a**, for different values of μ and $k^t/k^n = 0.8$.

well in this range, one could say that elasticity is enhanced by friction.

The sources of deviation from linearity are loss of contact and sliding induced by the application of the external force (both being ‘rearrangements’). The resulting changes in the contact network are shown in Fig. 2. For a sufficiently small force (not shown), the contact network is unchanged, and the response is fully elastic. As the force is increased, horizontal contacts are opened in a teardrop-shaped region below the point of application of the force, whose size increases with the force and decreases with friction (and with the stiffness of the particles, but the size of the region appears to saturate as the normal spring constant is increased above about $k^n = 2,000\bar{m}g/\bar{R}$ —where $\bar{R}$ and $\bar{m}$ are the mean particle radius and mass and g is the acceleration due to gravity—which is smaller than the value used in the simulations presented here). In the frictionless case, a region with no horizontal contacts comprises an ‘extremely’ anisotropic elastic system, in the sense that it corresponds locally to the hyperbolic limit of the elastic equations²⁰, related to the appearance of a zero-energy strain mode (note that in ref. 20, strict hyperbolic behaviour is shown to occur when all horizontal contacts are open). Two peaks are expected for sufficiently large anisotropy in an elastic system^{17,20}; in the hyperbolic limit, two delta peaks are obtained (this is not a result of the existence of force chains, which can be present in macroscopically isotropic systems¹⁷). When the ‘teardrop’ is sufficiently far from the floor (Fig. 2a, b) the pressure distribution is single-peaked, because the distortion induced by the force can be considered to be localized. Otherwise, the anisotropy induced by the external force reaches the floor (Fig. 2c), inducing a two-peaked response. The dependence of the size of the affected region on the depth of the slab was examined by simulating systems with different numbers of particles; for a given applied force, this size (measured in particle diameters) was not strongly sensitive to the depth of the slab (it actually decreased slightly with increasing system size). We may therefore conclude that, for identical external loads, very deep systems should exhibit single-peaked floor pressure distributions, whereas relatively shallow ones (as in some experiments^{8–12,16,18}) exhibit ‘hyperbolic-like’ two-peaked distributions. Because static friction (in the absence of sliding) is reversible, one can describe the stress distribution in either case by (incremental) isotropic or anisotropic, not necessarily homogeneous, elasticity. As mentioned, the ‘extreme’ limit of the anisotropic case (when all horizontal contacts are open) is of hyperbolic nature^{17,20}. We note that while the process in which particle contacts are opened cannot be described by elastic dynamics, it may be possible to describe the final state by elastic statics.

The localization of open contacts does not provide a full explanation for the results shown in Fig. 1, because for $\mu = 1$, the

response remains single-peaked (Fig. 1c) even when the teardrop reaches the floor (Fig. 2f). This turns out to be yet another consequence of friction. When μ is sufficiently large (and F_{ext} sufficiently small), its value is essentially immaterial, because sliding is never reached (of course, for sufficiently large forcing, even systems with large μ succumb, and a two-peaked response is obtained; this transition will not be described here). In this case,

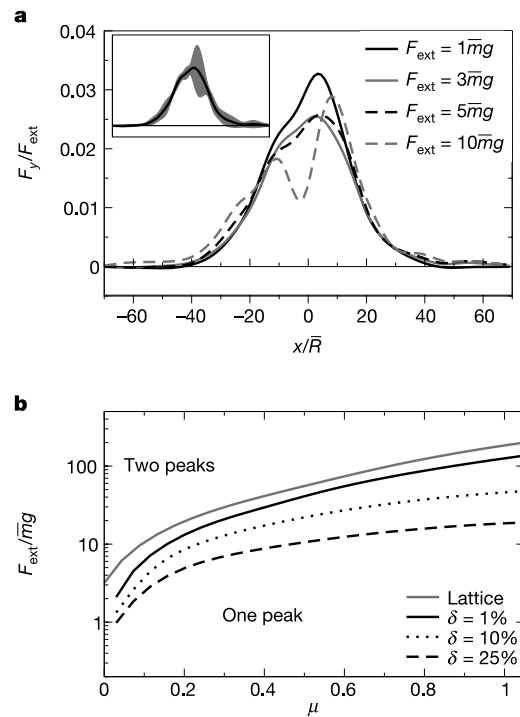


Figure 4 Effects of disorder and a phase diagram. **a**, The response of polydisperse systems, smoothed (see text) and averaged over five realizations of the disorder, with polydispersity $\delta = 0.25$ and $\mu = 0.2$, for various F_{ext} . The response is not symmetrical; however, the probability of a mirror image of a given configuration is the same as the original one, so we could include the mirror configurations to obtain a symmetrical response. It is interesting that an average over only five realizations (with smoothing) is sufficient to demonstrate the crossover from a single peak to a double peak, and even result (in some cases) in nearly symmetrical graphs. The inset indicates the root-mean-square fluctuations for $F_{\text{ext}} = 3\bar{m}g$. When far from the crossover force, the typical response of a realization resembles that of the ensemble average, whereas close to the crossover, stronger fluctuations are observed in the response. **b**, Schematic phase diagram (in the μ – F_{ext} plane) for the crossover from a single-peaked to a two-peaked response, for different degrees of polydispersity δ .

the ratio of tangential to normal stiffness, k^t/k^n , determines the response (Fig. 3a), which becomes nearly universal for large k^t/k^n . A spring network model (Fig. 3b) with $\mu = \infty$ (similar to the models studied in refs 21–25) can be used to estimate the elastic moduli in the teardrop domain. Using a criterion proposed in ref. 20, a crossover from a single peak to two peaks is predicted at $k^t/k^n \approx 0.3$, in agreement with our numerical findings (Fig. 3a). We note that, in practice, the expected value of k^t/k^n for realistic granular matter should satisfy $2/3 < k^t/k^n < 1$ (this follows from the Cattaneo–Mindlin model²⁶ for the contact of elastic spheres), which is well above the threshold for obtaining one peak. Here, unless otherwise noted, we use $k^t/k^n = 0.8$. The reduced anisotropy (for sufficiently large k^t/k^n) is due to the horizontal projections of the static frictional forces, which compensate for the lack of normal contact forces, and can prevent the opening of horizontal contacts²⁷, as indicated by the extended linear range in the presence of friction (Fig. 1d). In other words, static friction acts to retain an ‘effective connectivity’ between the grains when the horizontal contacts are open (as inside the teardrop), and render the system more isotropic than one would have anticipated in the absence of friction (but not exactly isotropic). Hence, for sufficiently large μ (and realistic k^t/k^n), the response is single-peaked. On the other hand, when μ is small (or F_{ext} large), sliding occurs, which increases the anisotropy and leads to a crossover to a two-peaked response. This crossover is shown as a function of F_{ext} (for $\mu = 0.2$) in Fig. 1b, and as a function of μ (for $F_{\text{ext}} = 15mg$) in Fig. 3c.

Polydisperse systems (with a uniform distribution of radii in the interval $[R - \delta R, R]$), where δ is a measure of the polydispersity, and R denotes the maximal particle radius, exhibit qualitatively similar behaviour to that described above. The main difference is that the range of values of F_{ext} for which the response is linear in F_{ext} , as well as the range of F_{ext} for which a single peak is obtained, decreases with increasing polydispersity, see Fig. 4a. Because the disorder in these systems induces fluctuations (which are not present in ordered systems), the results presented in Fig. 4a were smoothed by a convolution with a gaussian in the horizontal (x) direction: $\frac{1}{\sqrt{\pi}w} e^{-(x/w)^2}$ with $w = 6R$, and averaged over several realizations of the disorder (typically five).

Our results on the effect of F_{ext} , μ and the degree of polydispersity on the crossover are summarized in a schematic phase diagram, Fig. 4b. The diagram enables us to interpret several recent experimental studies of the response of granular materials to a localized force^{8–12,16,18}; different experiments correspond to different regions in this diagram. A single-peaked response was observed in experiments on disordered three-dimensional (3D) packings^{15,17} with small applied forces (a few times the particle weight), whereas multiple peaks were observed in experiments on ordered 3D packings¹⁰, where forces of a few thousand times the particle weight were applied. These peaks faded away for deeper systems. Experiments in 2D systems using photoelastic particles typically use a rather large applied force ($F_{\text{ext}} \approx 150mg$ in refs 8, 9), required to obtain a significant photoelastic effect. The interparticle forces in the nominally ordered case are reproduced qualitatively even with frictionless particles¹⁷; however, the introduction of friction improves this agreement to an essentially quantitative level. □

Received 8 August 2004; accepted 21 February 2005; doi:10.1038/nature03497.

- Nedderman, R. M. *Statics and Kinematics of Granular Materials* (Cambridge Univ. Press, Cambridge, 1992).
- Savage, S. B. in *Physics of Dry Granular Media* (eds Herrmann, H. J., Hovi, J. P. & Luding, S.) 25–95 (NATO ASI Series, Kluwer, Dordrecht, 1998).
- Jackson, R. in *Theory of Dispersed Multiphase Flow* (ed. Meyer, R. E.) 291–337 (Academic, New York, 1983).
- Wittmer, J. P., Claudin, P., Cates, M. E. & Bouchaud, J.-P. An explanation for the central stress minimum in sand piles. *Nature* **382**, 336–338 (1996).
- Bouchaud, J.-P., Claudin, P., Levine, D. & Otto, M. Force chain splitting in granular materials: a mechanism for large-scale pseudo-elastic behaviour. *Eur. Phys. J. E* **4**, 451–457 (2001).
- Tkachenko, A. V. & Witten, T. A. Stress propagation through frictionless granular material. *Phys. Rev. E* **60**, 687–696 (1999).

- Drescher, A. & de Josselin de Jong, G. Photoelastic verification of a mechanical model for the flow of a granular material. *J. Mech. Phys. Solids* **20**, 337–351 (1972).
- Geng, J. *et al.* Footprints in sand: the response of a granular material to local perturbations. *Phys. Rev. Lett.* **87**, 035506 (2001).
- Geng, J., Reydellet, G., Clément, E. & Behringer, R. P. Green's function measurements of force transmission in 2D granular materials. *Physica D* **182**, 274–303 (2003).
- Mueggenburg, N. W., Jaeger, H. M. & Nagel, S. R. Stress transmission through three-dimensional ordered granular arrays. *Phys. Rev. E* **66**, 031304 (2002).
- Moukarzel, C. F., Pacheco-Martinez, H., Ruiz-Suarez, J. C. & Vidales, A. M. Static response in disk packings. *Granular Matter* **6**, 61–66 (2004).
- da Silva, M. & Rajchenbach, J. Stress transmission through a model system of cohesionless elastic grains. *Nature* **406**, 708–710 (2000).
- Smid, J. & Novosod, J. Pressure distribution under heaped bulk solids. In *Proc. 1981 Powtech Conf., Institution of Chemical Engineers Symp.* **63**, D3/V/1–12 (1981).
- Brockbank, R., Huntley, J. M. & Ball, R. C. Contact force distribution beneath a three-dimensional granular pile. *J. Phys. II* **7**, 1521–1532 (1997).
- Vanel, L., Howell, D., Clark, D., Behringer, R. P. & Clément, E. Memories in sand: experimental tests of construction history on stress distributions under sandpiles. *Phys. Rev. E* **60**, R5040–R5043 (1999).
- Serero, D., Reydellet, G., Claudin, P., Clément, E. & Levine, D. Stress response function of a granular layer: quantitative comparison between experiments and isotropic elasticity. *Eur. Phys. J. E* **6**, 169–179 (2001).
- Goldenberg, C. & Goldhirsch, I. Force chains, microelasticity, and macroelasticity. *Phys. Rev. Lett.* **89**, 084302 (2002).
- Reydellet, G. & Clément, E. Green's function probe of a static granular piling. *Phys. Rev. Lett.* **86**, 3308–3311 (2001).
- Cundall, P. A. & Strack, O. D. L. A discrete numerical model for granular assemblies. *Geotechnique* **29**, 47–65 (1979).
- Otto, M., Bouchaud, J.-P., Claudin, P. & Socolar, J. E. S. Anisotropy in granular media: classical elasticity and directed-force chain network. *Phys. Rev. E* **67**, 031302 (2003).
- Duffy, J. & Mindlin, R. D. Stress-strain relations and vibrations of a granular medium. *J. Appl. Mech.* **24**, 585–593 (1957).
- Bathurst, R. J. & Rothenburg, L. Micromechanical aspects of isotropic granular assemblies with linear contact interactions. *J. Appl. Mech.* **55**, 17–23 (1988).
- Chang, C. S. & Ma, L. Elastic material constants for isotropic granular solids with particle rotation. *Int. J. Solids Struct.* **29**, 1001–1018 (1992).
- Gay, C. & da Silva, R. Anisotropic elastic theory of preloaded granular media. *Europhys. Lett.* **68**, 51–57 (2004).
- da Silva, R., Vidalenc, G. & Gay, C. Stress propagation in two dimensional frictional granular matter. Preprint at (<http://arXiv.org/cond-mat/0208214>) (2002).
- Johnson, K. L. *Contact Mechanics* 220 (Cambridge Univ. Press, Cambridge, 1985).
- Tournat, V. Probing weak forces in granular media through nonlinear dynamic dilatancy: clapping contacts and polarization anisotropy. *Phys. Rev. Lett.* **92**, 085502 (2004).

Acknowledgements We thank A. P. F. Atman, R. P. Behringer, P. Claudin, E. Clément, J. Geng, N. Mueggenburg, M. van Hecke, W. van Saarloos and T. A. Witten for discussions. This work was supported by the Israel Science Foundation (ISF) and the US-Israel Binational Science Foundation (BSF).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.G. (isaac@eng.tau.ac.il).

Nonlinear elasticity in biological gels

Cornelis Storm^{1*}, Jennifer J. Pastore², F. C. MacKintosh³, T. C. Lubensky^{1,2} & Paul A. Janmey^{1,2}

¹Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, Pennsylvania 19104, USA

²Institute for Medicine and Engineering, University of Pennsylvania, 3340 Smith Walk, Philadelphia, Pennsylvania 19104, USA

³Division of Physics and Astronomy, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

* Present address: Instituut-Lorentz for Theoretical Physics, LION, Universiteit Leiden, PO Box 9506, NL-2300RA Leiden, The Netherlands

The mechanical properties of soft biological tissues are essential to their physiological function and cannot easily be duplicated by synthetic materials. Unlike simple polymer gels, many biological materials—including blood vessels¹, mesentery tissue², lung parenchyma³, cornea⁴ and blood clots⁵—stiffen as they are strained, thereby preventing large deformations that could threaten tissue integrity. The molecular structures and design

principles responsible for this nonlinear elasticity are unknown. Here we report a molecular theory that accounts for strain-stiffening in a range of molecularly distinct gels formed from cytoskeletal and extracellular proteins and that reveals universal stress-strain relations at low to intermediate strains. The input to this theory is the force-extension curve for individual semiflexible filaments and the assumptions that biological networks composed of these filaments are homogeneous, isotropic, and that they strain uniformly. This theory shows that systems of filamentous proteins arranged in an open crosslinked mesh invariably stiffen at low strains without requiring a specific architecture or multiple elements with different intrinsic stiffness.

The filamentous architectures of two representative strain-stiffening networks—namely, neuronal intermediate filaments (the major cytoskeletal element of axons) and fibrin protofibrils (the fundamental polymers forming blood clots)—are shown in Fig. 1. These networks have large solvent-filled spaces, and their filaments are only gently curved between points of intersection. Stiffness, quantified by the shear moduli of a variety of biopolymer networks, is plotted as a function of applied strain in Fig. 2. Other studies have reported similar strain-stiffening in crosslinked actin^{6,7}, keratin⁸ and cytoplasmic gels⁶. The systems that we consider vary over orders of magnitude in stiffness and in the strains they tolerate: neurofilament networks can be deformed over 400% before failing, whereas actin networks rupture at strains of 20%.

The response of a single elastic filament to an applied force has been the subject of much study^{9–15}. This force-response is dominated by entropy: because there are many curled-up configurations and only one that is perfectly straight, stretching a flexible filament reduces its conformational entropy and thus produces an opposing force. The ensuing elastic behaviour of semiflexible polymers is inherently nonlinear, as even modest strains pull the constituent filaments nearly straight, in marked contrast with conventional rubbers¹⁶.

Polymer theory distinguishes three types of filaments, characterized by two length scales: the persistence length l_p (the typical length scale for the decay of tangent-tangent correlations) and the contour length L_c . A filament is considered flexible when $l_p \ll L_c$ and rigid when the opposite holds. Completely flexible filaments exhibit a purely entropic elastic response, whereas rigid filaments display no entropic elasticity. Most biologically relevant filaments are in a third intermediate category: semiflexible filaments with l_p and L_c of comparable magnitude. These filaments do not form loops and knots, yet they are sufficiently flexible to have significant thermal bending fluctuations. In networks, the relevant contour length is the distance between network junction points or crosslinks, which we also call L_c below.

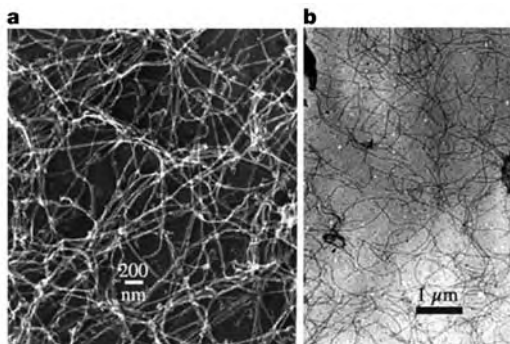


Figure 1 Neurofilament and fibrin protofibril networks. These TEM images show the finite excess of filament contour length between crosslinks and overlap points. **a**, Metal-shadowed neurofilaments, and **b**, uranyl acetate-stained fibrin protofibrils, prepared as described in refs 25 and 26, respectively.

A theory for the force-response of such semiflexible filaments and the corresponding network shear modulus has been proposed¹². This model is based on an energy functional of the form

$$E = \int_0^{L_c} dz \left\{ \frac{\kappa}{2} |\mathbf{u}''|^2 + \frac{f}{2} |\mathbf{u}'|^2 \right\} \quad (1)$$

where z is the projected coordinate along the end-to-end vector, κ is the bending stiffness (related to the persistence length as $\kappa = k_B T l_p$) and f is the applied force. $\mathbf{u}(z)$ describes the deviation of the filament from its straight conformation. The length $L = L(f; L_c)$ of the end-to-end vector is calculated to harmonic order in $\mathbf{u}(z)$ using the equipartition theorem¹⁷ and the geometric relation between contour length and $\mathbf{u}(z)$. The result is most conveniently expressed in terms of the scaled difference between the extension at force f and that at zero force

$$\delta \tilde{L} = \frac{L(f; L_c) - L(0; L_c)}{L_c^2 / l_p} = \frac{1}{\pi^2} \sum_{n=1}^{\infty} \frac{\varphi}{n^2(n^2 + \varphi)} \quad (2)$$

where $\varphi = f L_c^2 / \kappa \pi^2$ is a dimensionless force (see the inset to Fig. 3). The scaling force $\kappa \pi^2 / L_c^2$ is the threshold force for the Euler buckling instability in thin cylinders^{18,19}. Equation (2) can be inverted to yield a force-extension relation that, like the worm-like chain model applied to DNA^{9–11}, diverges as $f \sim (L - L_c)^{-2}$ as $L \rightarrow L_c$.

To get from the force-extension curve of single filaments to the

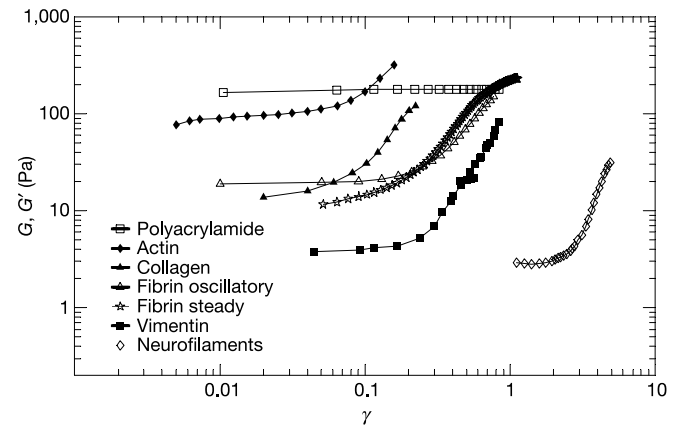


Figure 2 Dynamic shear storage moduli measured at different strain amplitudes for a series of crosslinked biopolymer networks. The real part, G' , of the storage modulus reduces to the shear modulus G at zero frequency. Data shown are G' (at 10 rad s⁻¹) values for F-actin, fibrin, collagen, vimentin and polyacrylamide, and shear modulus G for fibrin and neurofilaments, plotted as a function of the dimensionless strain γ . Dynamic shear moduli were measured in a strain-controlled rheometer (RFS-II, Rheometrics), which applies a sinusoidally varying strain with controllable maximal strain amplitude and computes the elastic storage moduli from the amplitude and phase shift of the resulting stress. Data analysis in the proprietary software assumes that the stress response is also a sinusoidal function, but in the nonlinear range of strains this relationship is not strictly true and the resulting stress function deviates from a sinusoid by having a sharper peak at maximal stress. To determine if this complication perturbs the reported strain-stiffening results, we also computed G by measuring stress at constantly increasing strains by using the steady rate function of the RFS-II, which imposes a constant increase in strain (5% s⁻¹) and reports the resulting stress at very small strain increments (<0.1%). The steady strain data overlap those from an oscillatory measurement of G' for the same sample of fibrin, verifying that oscillatory measurements accurately report the magnitude of strain stiffening under the conditions of our experiments. Sample preparation: F-actin/filamin gels were prepared as described²⁷; vimentin (2 mg ml⁻¹) was a gift of P. Traub and prepared as described²⁸; neurofilaments (3 mg ml⁻¹) were prepared as described²³; rat tail collagen (2 mg ml⁻¹) was obtained from Sigma; polyacrylamide/bisacrylamide (5%) was polymerized with ammonium persulphate and TEMED by standard methods; fibrinogen was purified from salmon blood plasma as described²⁹ and dialysed into 50 mM Tris, 450 mM NaCl, pH 8.5, under which conditions only fibrin protofibrils form and lateral association into thicker bundles is prevented.

bulk elastic properties of filament networks requires a model for the network geometry. We consider a model isotropic network in which pairs of crosslink points (nodes) are connected by links composed of independent semi-flexible filaments. We assume that no torques are exerted at nodes so that filaments stretch or compress but do not bend in response to forces. We also assume that the network is sufficiently random that the filaments are isotropically distributed in the absence of stress. We allow for a distribution of end-to-end separations between nodes, with an average equal to $L(f=0;L_c)$, where $L(f=0;L_c) = L_c(1 - L_c/6l_p)$ is the equilibrium end-to-end length of a polymer segment with a total contour length L_c . In response to external stresses, network nodes will displace and filaments will stretch or compress, that is, the network will strain. Under strain, the position $\mathbf{R}_{0\alpha}$ of node α in the unstrained network will transform to a new position $\mathbf{R}_\alpha = \Lambda\alpha\mathbf{R}_{0\alpha}$ where $\Lambda\alpha$ is the local (not necessarily symmetric) Cauchy deformation tensor. In random networks such as those we consider, $\Lambda\alpha$ will in general depend on α , that is, deformations will be non-affine. We make the simplifying assumption that deformations are affine, and take $\Lambda\alpha = \Lambda$ independent of α . Although there is no a priori reason to believe the affine approximation is valid, recent theoretical and experimental studies suggest that it is a good approximation for densely cross-linked filaments of high molecular weight^{20–22}. As all forces are transmitted by links, and the force on each link is determined by its length, we can calculate the nonlinear stress tensor by averaging the force per area exerted by a single link over all magnitudes and directions of $\mathbf{r}$. The result is

$$\sigma_{ij}^T = \frac{\rho}{\det\Lambda} \left\langle f(|\Lambda\mathbf{r}|) \frac{(\Lambda_{il}r_l)(\Lambda_{jk}r_k)}{|\Lambda\mathbf{r}|} \right\rangle_{P(\mathbf{r})}$$

where ρ is the number of links per unit volume and the average is over the probability distribution $P(\mathbf{r})$ for separations $\mathbf{r}$. This stress tensor measures the force per unit area of the deformed sample and is symmetric.

We consider only the volume-conserving simple shear deformation produced in standard rheometers characterized by:

$$\Lambda = \begin{pmatrix} 1 & 0 & \gamma \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

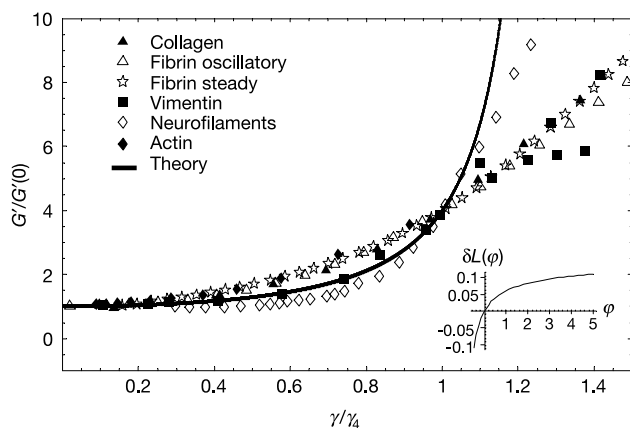


Figure 3 Scaled modulus–strain curves for various biopolymer networks compared to theory and the scaled force–extension relation for a semiflexible polymer. Main panel, data of Fig. 2 scaled as described in the text (symbols) and theory (solid line). The theoretical curve was obtained assuming a uniform distribution of filament end-to-end lengths (that is, all equal to the mesh size), and averaging over all orientations in three dimensions. Inset, the dimensionless force versus extension curve described by equation (2).

In this case, the shear modulus is related to the xz -component of the stress tensor via $G(\gamma) = \sigma_{xz}(\gamma)/\gamma$.

To evaluate the shear modulus, we specify the bond probability distribution and force law $f(L)$. The simplest model is one in which $f(L)$ is determined by equation (2) and the distribution function sets the distance between all nodes equal to the relaxed end-to-end length $L(f=0;L_c)$ of the links, all of which we assume have the same contour length L_c . In this model, no new lengths are introduced by the averaging process, and the curve of shear modulus versus strain exhibits a universal scaling similar to that of the force–extension curve onto which data described by it will collapse. Figure 3 plots experimentally measured shear moduli G scaled by their value G_0 at zero strain versus strain γ normalized by the strain γ_4 at which $G = 4G(0)$. The value of 4 is arbitrarily chosen as the degree of stiffening that most systems examined exhibit before failure. The data collapse approximately onto the universal curve at modest strains, suggesting that the initial strain-stiffening in biopolymer networks is dominated by entropic effects that are well captured by our theory.

The theory developed thus far concerns only the low strain regime where entropic elasticity dominates. Figure 3 shows that at high strains departures from the theory become apparent, presumably because entropy alone does not capture all the physics of these systems.

We now modify the model to address these deviations, assuming that they arise from enthalpic contributions to $f(L)$, which we describe by introducing a stretch modulus K , measuring the longitudinal compliance of our previously inextensible filaments ($K = \infty$). The force–extension relation, equation (2), for finite K becomes

$$L_K(f;L_c) = \left(1 + \frac{f}{K}\right)L\left(f\left(1 + \frac{f}{K}\right);L_c\right)$$

where $L(f;L_c)$ is defined by equation (2). The persistence length and the stretch modulus of a rod are not independent: for a cylindrical tube of radius r they are related through²³ $l_p/K = r^2/4k_B T$.

The restriction that all filaments have the same end-to-end length is also not realistic, even if they have the same contour length. Crosslinking of thermally undulating filaments creates local pairwise node separations that differ from the zero-force end-to-end lengths of the filaments. By generalizing the classical theory of elasticity of crosslinked flexible polymers, we take the distribution function of end-to-end lengths of filaments in our network to be the equilibrium distribution of end-to-end lengths of semiflexible inextensible filaments of contour length L_c (ref. 24):

$$P(r) = \frac{2l_p}{L_c^2} \sum_{n=1}^{\infty} (\pi n)^2 (-1)^{n+1} \exp\left[-\frac{l_p(\pi n)^2}{L_c} \left(1 - \frac{r}{L_c}\right)\right]$$

An important consequence of having distributed lengths in this manner is that the network's filaments no longer individually experience zero force (as they do for a distribution with all lengths equal to the mean): some are stretched while others are compressed relative to their relaxed length. As a consequence, bulk mechanical equilibrium at zero isotropic stress will occur not at zero strain but at an isotropic strain described by a deformation tensor $\Lambda_{0,ij} = \Lambda_0\delta_{ij}$. The deformation Λ_0 is determined by the condition that the isotropic part of the stress tensor $\sigma^I\delta_{ij}$ vanishes. The angular average in the expression for the stress tensor in the presence of isotropic deformations is calculated to yield

$$\sigma^I(\Lambda_0) = \frac{\rho}{3\Lambda_0^2} \int_0^{L_c} dr \, r \, P(r) f(\Lambda_0 L(f=0;L_c)) = 0$$

as the condition determining Λ_0 . This equation is solved numerically for particular parameter values. Λ_0 is not an extra parameter—it is determined by the force–extension relation and the distribution function, both of which are functions of the persistence length, the

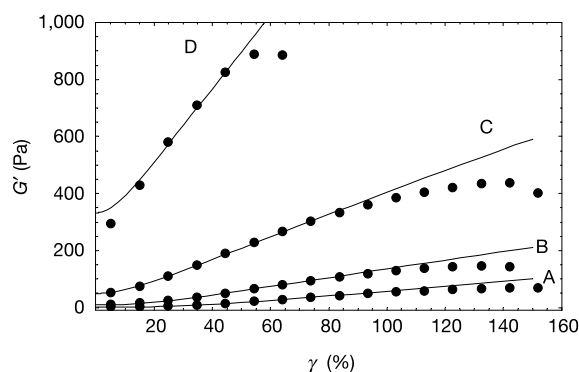


Figure 4 Experimental data for fibrin protofilaments (dots) at various concentrations, and corresponding theoretical curves (solid lines) as computed from the extended theory including a stretch modulus. Best-fit values were determined for A–D as follows: A, $c = 0.5 \text{ mg ml}^{-1}$ ($\xi = 0.39 \mu\text{m}$), $K = 67 \text{ pN}$, $\Lambda_0 = 0.948$; B, $c = 1.0 \text{ mg ml}^{-1}$ ($\xi = 0.27 \mu\text{m}$), $K = 58 \text{ pN}$, $\Lambda_0 = 0.969$; C, $c = 2.0 \text{ mg ml}^{-1}$ ($\xi = 0.19 \mu\text{m}$), $K = 73 \text{ pN}$, $\Lambda_0 = 0.981$; D, $c = 4.5 \text{ mg ml}^{-1}$ ($\xi = 0.12 \mu\text{m}$), $K = 110 \text{ pN}$, $\Lambda_0 = 0.991$. See text for definitions of parameters.

contour length and the stretch modulus only.

The extended theory makes specific predictions about how strain-stiffening scales with network density, and we test this by fitting to a series of modulus-strain curves for fibrin at different mass concentrations. Figure 4 shows the experimental data and the theoretical curves. The persistence length was experimentally determined to be $0.5 \mu\text{m}$ by computing the length scale for the decay of tangent–tangent correlations based on digitized microscopic images of fibrin filaments such as those in Fig. 1. The mesh size ξ of an isotropic network of polymers is calculated from the mass per unit length of polymer, λ , and the mass density, c . The number of links per unit volume is $\rho = \lambda/cL_c = 3/(\xi^2 L_c)$. We assume that $\xi \approx L_c$ so that $\rho = 3/\xi^3$. As ρ , ξ and Λ_0 are all determined by the mass concentration, our model has only one free parameter, the stretch modulus K . The best-fit values for K from Fig. 4 are of the order of 50–100 pN, very close to the elastic-rod estimate of $K = 4k_B T l_p/r^2$ for a rod of radius $r = 10 \text{ nm}$ and $l_p = 0.5 \mu\text{m}$.

For all parameter values considered, Λ_0 is smaller than, but of the order of, one, implying a small isotropic compression due to the asymmetric character of the single filament force–extension relation (see the inset to Fig. 3). As force rises steeply for extensions but remains relatively flat for compressions, stretched filaments contribute more to the residual stress after crosslinking, and the network as a whole will shrink.

Nonlinear elasticity, and specifically strain-stiffening, is generic to any network composed of semiflexible filamentous proteins. The degree of strain-stiffening can exceed tenfold increases in shear moduli under strains as small as 20%. The strain at which stiffening becomes significant depends strongly on the persistence length of the filament. Stiffer filaments, like F-actin or collagen, stiffen at a few per cent strain whereas more flexible filaments like vimentin stiffen only at larger strains, approaching 100%. The functional form of strain-stiffening at intermediate strain is generic for all semiflexible systems, but the maximal degree of strain-stiffening depends on such molecular details as the compliance of filaments to extension. Likewise, the strain at which the networks rupture also depends on the nature of the chemical bonds holding the filaments together.

These results have implications for cytoskeletal and other biopolymer networks. They suggest that the design of artificial biomaterials to replace tissues such as the arterial wall, whose function requires nonlinear elastic response¹, could be facilitated by using polymers of stiffness and mesh size appropriate to produce strain-stiffening at the required range of deformations. Using this nonlinear passive elasticity, biological systems like the cytoskeleton

can actively manipulate their stiffness by local contraction of the network using motor proteins.

The inherent strain-stiffening of semiflexible polymer networks is only a starting point for the designs that have evolved to endow biological materials with their mechanical properties. Many strain-stiffening tissues have strikingly ordered stiff filaments co-assembled with an amorphous matrix. Variation of ordered elements in harder material like bone or wood provides other mechanisms to create nonlinear elasticity. The significance of the present work is that it shows such geometrical ordering is not required to produce the high degree of strain-stiffening that is an essential aspect of the elasticity of isotropic networks of crosslinked semiflexible polymers. □

Received 27 May 2004; accepted 7 March 2005; doi:10.1038/nature03521.

- Shadwick, R. E. Mechanical design in arteries. *J. Exp. Biol.* **202**, 3305–3313 (1999).
- Fung, Y. A *First Course in Continuum Mechanics* (Prentice Hall, Englewood Cliffs, 1994).
- Karakaplan, A. D., Bieniek, M. P. & Skalak, R. A mathematical model of lung parenchyma. *J. Biomech. Eng.* **102**, 124–136 (1980).
- Hjortdal, J. O. Extensibility of the normo-hydrated human cornea. *Acta Ophthalmol. Scand.* **73**, 12–17 (1995).
- Shah, J. V. & Janmey, P. A. Strain hardening of fibrin gels and plasma clots. *Rheologica Acta* **36**, 262–268 (1997).
- Janmey, P. A. *et al.* The mechanical properties of actin gels. Elastic modulus and filament motions. *J. Biol. Chem.* **269**, 32503–32513 (1994).
- Xu, J., Tseng, Y. & Wirtz, D. Strain hardening of actin filament networks. Regulation by the dynamic cross-linking protein alpha-actinin. *J. Biol. Chem.* **275**, 35886–35892 (2000).
- Ma, L., Xu, J., Coulombe, P. A. & Wirtz, D. Keratin filament suspensions show unique micromechanical properties. *J. Biol. Chem.* **274**, 19145–19151 (1999).
- Fixman, M. & Kovac, J. Polymer conformational statistics. III. Modified Gaussian models of stiff chains. *J. Chem. Phys.* **58**, 1564–1568 (1973).
- Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurement of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
- Marko, J. & Siggia, E. Stretching DNA. *Macromolecules* **28**, 8759–8770 (1995).
- MacKintosh, F., Käs, J. & Janmey, P. Elasticity of semiflexible biopolymer networks. *Phys. Rev. Lett.* **75**, 4425–4428 (1995).
- Kroy, K. & Frey, E. Force-extension relation and plateau modulus for wormlike chains. *Phys. Rev. Lett.* **77**, 306–309 (1996).
- Morse, D. Viscoelasticity of tightly entangled solutions of semiflexible polymers. *Phys. Rev. E* **58**, R1237–R1240 (1998).
- Gittes, F. & MacKintosh, F. C. Dynamic shear modulus of a semiflexible polymer network. *Phys. Rev. E* **58**, R1241–R1244 (1998).
- Treloar, L. R. G. *The Physics of Rubber Elasticity* (Clarendon, Oxford, 1975).
- Landau, L. D. & Lifshitz, E. M. *Statistical Physics I* (Pergamon, Oxford, 1980).
- Landau, L. D. & Lifshitz, E. M. *Theory of Elasticity* (Pergamon, Oxford, 1986).
- Gittes, F., Mickey, B., Nettleton, J. & Howard, J. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J. Cell Biol.* **120**, 923–934 (1993).
- Head, D. A., Levine, A. J. & MacKintosh, F. C. Deformation of cross-linked semiflexible polymer networks. *Phys. Rev. Lett.* **91**, 108102 (2003).
- Wilhelm, J. & Frey, E. Elasticity of stiff polymer networks. *Phys. Rev. Lett.* **91**, 108103 (2003).
- Gardel, M. L. *et al.* Elastic behavior of cross-linked and bundled actin networks. *Science* **304**, 1301–1305 (2004).
- Grosberg, A. Y. & Khokhlov, A. R. *Statistical Physics of Macromolecules* (American Institute of Physics Press, New York, 1994).
- Wilhelm, J. & Frey, E. Radial distribution function of semiflexible polymers. *Phys. Rev. Lett.* **77**, 2581–2584 (1996).
- Leterrier, J. F., Kas, J., Hartwig, J., Vegners, R. & Janmey, P. A. Mechanical effects of neurofilament cross-bridges. Modulation by phosphorylation, lipids, and interactions with F-actin. *J. Biol. Chem.* **271**, 15687–15694 (1996).
- Janmey, P. A. & Ferry, J. D. Gel formation by fibrin oligomers without addition of monomers. *Biopolymers* **25**, 1337–1344 (1986).
- Janmey, P. A., Hvidt, S., Lamb, J. & Stossel, T. P. Resemblance of actin-binding protein/actin gels to covalently crosslinked networks. *Nature* **345**, 89–92 (1990).
- Janmey, P. A., Euteneuer, U., Traub, P. & Schliwa, M. Viscoelastic properties of vimentin compared with other filamentous biopolymer networks. *J. Cell Biol.* **113**, 155–160 (1991).
- Wang, L. Z. *et al.* Purification of salmon clotting factors and their use as tissue sealants. *Thromb. Res.* **100**, 537–548 (2000).

Acknowledgements We are grateful to J.-F. Leterrier, S. Hvidt, P. Traub, J. Hartwig and E. Sawyer for collaboration in producing the protein networks and electron micrographs. This work was supported in part by the US-NIH and NSF/MRSEC programmes (P.A.J., T.C.L., C.S., J.J.P.) and by the Kavli Institute for Theoretical Physics at the University of California (E.C.M., P.A.J.), where some of this work was initiated.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.S. (cstorm@lorentz.leidenuniv.nl).

Plate-wide stress relaxation explains European Palaeocene basin inversions

Søren B. Nielsen, Erik Thomsen, David L. Hansen & Ole R. Clausen

Department of Earth Sciences, University of Aarhus, Finlandsgade 8, DK-8200 Aarhus N, Denmark

During Late Cretaceous and Cenozoic times, many Palaeozoic and Mesozoic rifts and basin structures in the interior of the European continent underwent several phases of inversion (the process of shortening a previously extensional basin)¹. The main phases occurred during the Late Cretaceous and Middle Palaeocene, and have been previously explained by pulses of compression, mainly from the Alpine orogen^{2–5}. Here we show that the main phases differed both in structural style and cause. The Cretaceous phase was characterized by narrow uplift zones,

reverse activation of faults, crustal shortening, and the formation of asymmetric marginal troughs. In contrast, the Middle Palaeocene phase was characterized by dome-like uplift of a wider area with only mild fault movements, and formation of more distal and shallow marginal troughs. A simple flexural model explains how domal, secondary inversion follows inevitably from primary, convergence-related inversion on relaxation of the in-plane tectonic stress. The onset of relaxation inversions was plate-wide and simultaneous, and may have been triggered by stress changes caused by elevation of the North Atlantic lithosphere by the Iceland plume⁶ or the drop in the north–south convergence rate between Africa and Europe⁷.

The sedimentary succession southwest of the Sorgenfrei-Tornquist Zone in the central part of the Danish Basin has recorded the two phases of inversion (Fig. 1). The Late Cretaceous phase occurred during a general increase of relative sea level, which provided space for a background thickness of chalk sediments. The inversion is visible in thinning of the sedimentary sequences onto the inversion ridge and the presence of internal unconformities and embedded sandstone bodies⁸, which show that the main

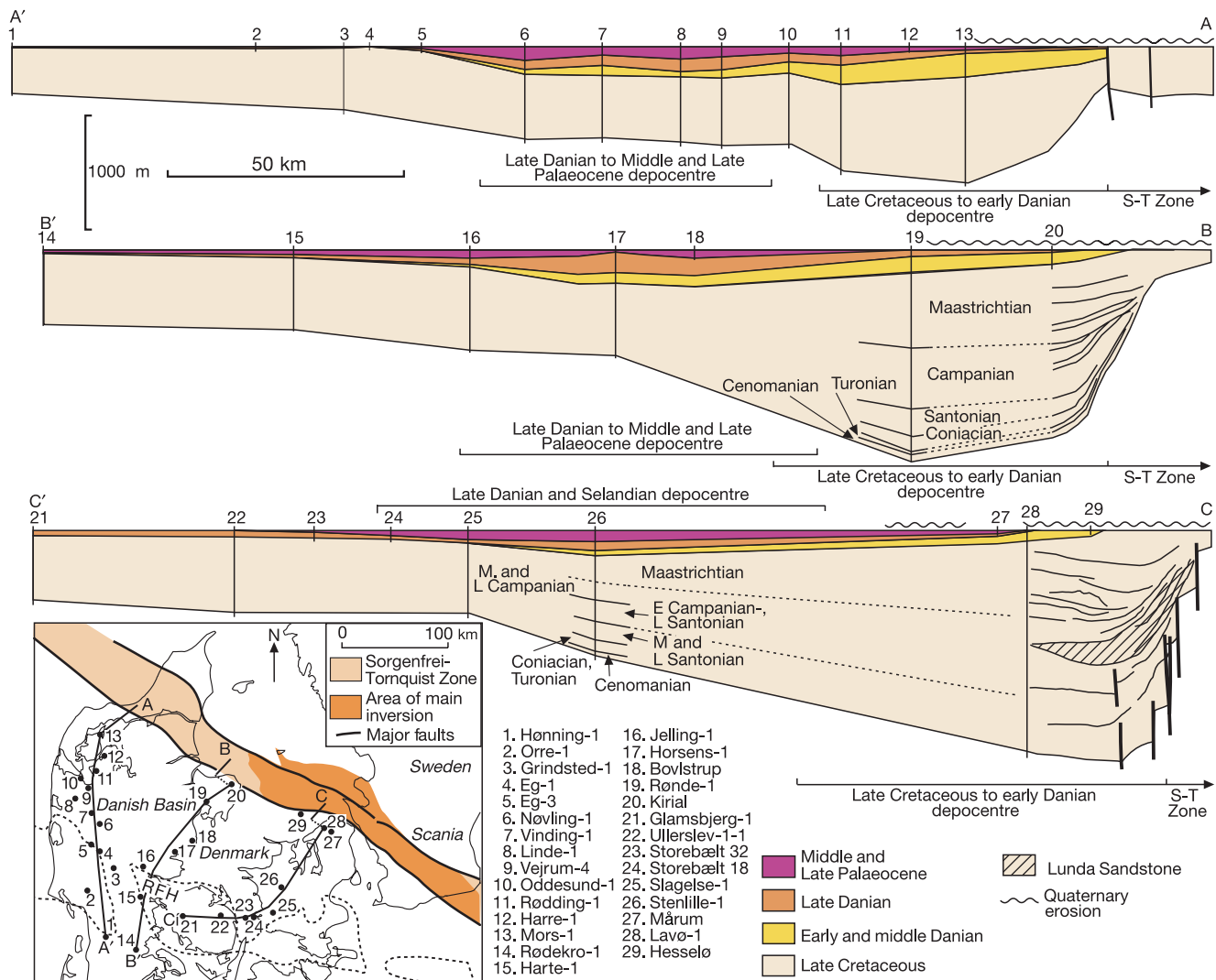


Figure 1 Geological profiles. Geosections across the marginal troughs of the Sorgenfrei-Tornquist Zone (STZ) representing the mildly inverted northwestern part (A–A' and B–B') and the strongly inverted southeastern part (C–C'). The Late Cretaceous marginal trough was formed during primary inversion, culminating during the Santonian and Campanian^{8,12}. The secondary, more distal, Palaeocene trough is shallower and more symmetric than the

Cretaceous trough. For section B–B' simple calculations based on the slope of the basal Upper Cretaceous and Palaeocene beds indicate that the Palaeocene uplift was only 10–15% of the total primary and secondary uplift. The datum level is the top of the Upper Palaeocene deposits in sections A–A' and B–B' and the top of the Middle Palaeocene deposits in section C–C'. R-FH, Ringkøbing-Fyn High; E, Early; M, Middle; L, Late.

uplift occurred during the Campanian and lower Maastrichtian. Simultaneously, the chalk basin southwest of the Sorgenfrei-Tornquist Zone developed a marked, asymmetric deepening towards the inversion axis (Fig. 1). This development conforms closely to the Late Cretaceous inversion history of other structures across the European continent³.

The structural style and compressional cause^{2,3} of the Late Cretaceous inversion phase agree with the results of model experiments^{9–11}. Compression and shortening (for example, 5–10%) of a model rift squeezes sediments out of the rift along the inversion axis because the crust thickens. The order of magnitude of erosion depth along the inversion axis is 10^3 m. Simultaneously, a positive lithospheric load develops along the central inversion axis (relative to the pre-compression state) because the eroded lighter material is replaced by underlying, denser material (crust and compacted sediments). The primary marginal troughs form as a consequence of flexural isostatic compensation of this load, enhanced by sediment loading from the troughs themselves^{9–11} (Figs 1, 2). The internal lithospheric load along the inversion axis prevails after erosion of all topography because rift inversion starts with an attenuated crystalline crust and a (thick) sedimentary cover. Thus, the flexural foredeeps of inversion zones and the loads that keep them in place are very resistant features.

The Middle Palaeocene inversion of the Sorgenfrei-Tornquist Zone, however, was entirely different. During the Danian, the depocentre southwest of the zone shifted away from the inversion zone, and at the beginning of the late Danian more distal, shallow and symmetric depocentres developed (Fig. 1). Slightly later ($\sim 10^5$ yr), during the early Selandian, the composition of the chalk component, constituting 50–70% of the sediment, changed from primarily biogenic to reworked Late Cretaceous chalk¹², marking the onset of the Middle Palaeocene inversion.

The more distal position and symmetric shape of the associated shallow, marginal troughs show that the Palaeocene inversion phase was more gentle and dome-like than the Late Cretaceous phase. We suggest that the cause was also different, as additional tectonic compression would simply continue the deepening of the Late Cretaceous asymmetric marginal troughs¹¹. In fact, our model experiments (Fig. 2) indicate that the Middle Palaeocene inversion of the Sorgenfrei-Tornquist Zone is best explained by relaxation of the in-plane tectonic stress rather than by compression.

The mechanism is well known¹³ and occurs because the equilibrium depth of the lithosphere flexure caused by the loads of the central inversion zone and the fill of the primary marginal troughs is sensitive to the level of in-plane stress. Thus, the Late Cretaceous compressional state of the lithosphere over-deepened the flexure slightly, and secondary relaxation inversion followed as an inevitable consequence of a relaxation of the in-plane compression in the Middle Palaeocene (Figs 1, 2). The surface expression of the relaxation inversion is a low-amplitude (order of magnitude 10^2 m), smooth, domal uplift of the central inversion zone and the proximal areas of the primary marginal troughs, and subsidence of more distal, shallow and symmetric secondary marginal troughs (Fig. 2). At the southern end of the Sorgenfrei-Tornquist Zone around Scania, the Middle Palaeocene inversion apparently was relatively narrow and partly detached from the marginal troughs⁸. We attribute this to a major discontinuity in the lithospheric structure in the area¹⁴.

The Middle Palaeocene inversion phase in northwest Europe has invariably been linked to additional compression^{2–5}. Relaxation inversion in the aftermath of compression was never considered. In the light of our findings for the Sorgenfrei-Tornquist Zone, we have examined the inversion history of several key European structures. It appears that the Palaeocene phase possesses all the characteristics of relaxation inversion: (1) the uplifts were domal, generally non-ruptural and mostly affected a wider area than the Late Cretaceous phases. (2) The marginal troughs were shallower and developed in a more distal position. (3) Where

erosion depths of the secondary phase are available, they are about 10% of the total erosion. Furthermore, the Middle Palaeocene inversions were simultaneous within the resolution of the stratigraphic data (a few Myr), in contrast to the ~ 20 Myr duration of the Late Cretaceous phase, indicating that they were triggered

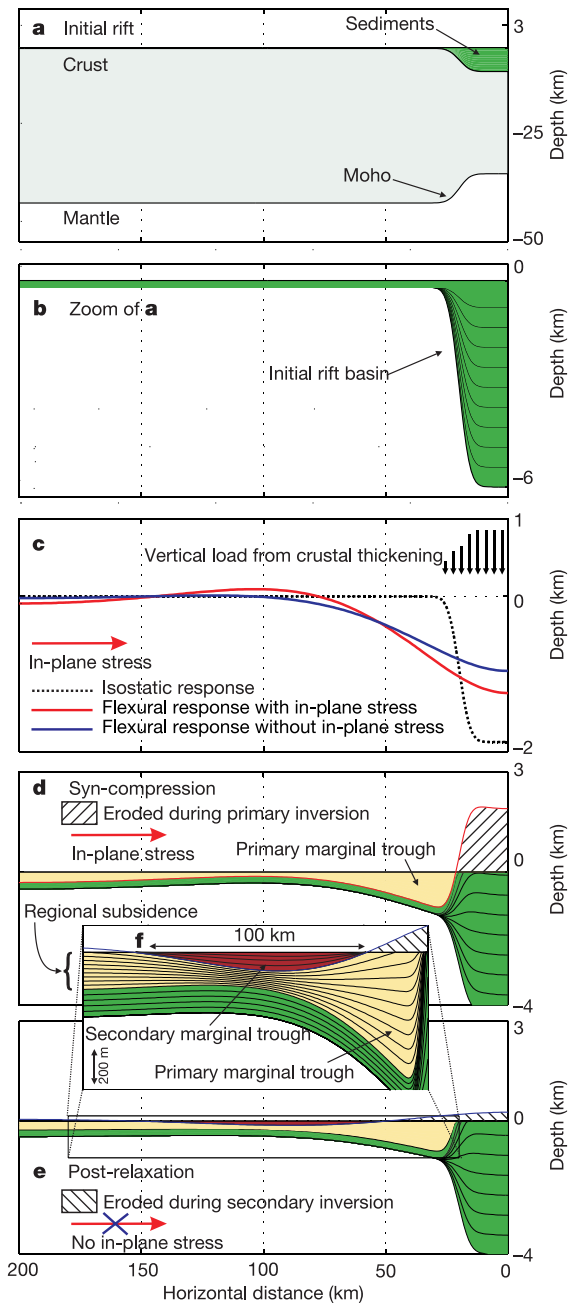


Figure 2 Mechanisms of primary and secondary inversion. **a, b**, Geometry of initial rift basin (rifting has ceased). **c**, Deflection of upper mantle caused by (blue) load of central inversion zone and (red) loading together with in-plane compression. **d**, Geometry of basin fill after primary inversion (in-plane compression). **e**, Geometry after secondary inversion (no in-plane compression). **f**, Enlarged view of the marginal troughs after secondary inversion. The amplitude and wavelength of secondary inversion depends on the flexural rigidity of the lithosphere, the magnitude of the load, density contrasts in the lithosphere and the size of stress change involved. Loading across the structure by differentially compacting sediments and ductile stress relaxation may enhance the amplitudes of vertical displacements. Typically, the erosion depth of the secondary phase amounts to about 10% of the total erosion. Model densities are $2,400 \text{ kg m}^{-3}$ for sediments, $2,900 \text{ kg m}^{-3}$ for crust, and $3,300 \text{ kg m}^{-3}$ for lithospheric mantle. The bending stiffness of the elastic plate is $2 \times 10^{21} \text{ N m}^{-1}$. The magnitude of in-plane stress is $5 \times 10^{12} \text{ N m}^{-1}$.

by a single event causing a plate-wide stress change.

The Tornquist-Teisseyre Zone across Poland experienced a strong Late Cretaceous compressional inversion, with an estimated 2–3 km of erosion¹⁵. In analogy with the Sorgenfrei-Tornquist Zone, we interpret the narrow Cretaceous depocentres as the remains of the primary marginal troughs, and the wider and more distal Palaeocene depocentre as the secondary marginal troughs (Fig. 3a). The elongate area of uplift separating the depocentres is, in accordance with model experiments (Fig. 2), wider than the ridge formed during the primary inversion¹⁶.

The Danish Central Graben saw episodic and relatively weak Late Cretaceous, fault-controlled inversion of Late Jurassic–Early Cretaceous depocentres¹⁷. Post-Danian¹⁷, possibly during the Middle Palaeocene¹⁸, the structural style changed to gentle folding and doming of a wider area^{17,18}. The Dutch Central Graben¹⁹ experienced a stronger Late Cretaceous inversion with patchy removal of the chalk sequence (Fig. 3a), followed by differential post-Danian subsidence across the inverted graben¹⁹. In both cases, the Middle Palaeocene change of structural style and the influence of the gentle differential vertical movements on the post-Danian Palaeocene sediments (Fig. 3a) can be explained by relaxation inversion.

The Lower Saxony Basin (Fig. 3a) experienced strong Late Cretaceous inversion with several kilometres of erosion and formation of pronounced marginal troughs²⁰, which signal the existence of a loading-induced lithosphere flexure. This primary inversion mode was followed by domal, non-ruptural uplift during the Palaeocene^{20,21}, which we identify with secondary relaxation inversion. Evidence of the change in inversion style comprises deep truncation of the deposits of the primary marginal troughs along the margins of the inversion zone and relaxation movements along former Late Cretaceous reverse faults dated by Palaeocene sediments²¹. The proximal part of the post-Danian Palaeocene depocentre to the north (Fig. 3a) is in the right position to be a secondary marginal

trough, but detailed thickness information is not available to us.

The inverted²² Weald-Boulonnais area (Fig. 3a) is flanked by shallow, symmetric depocentres of early Thanetian (NP6) age²³, which have all the characteristics of secondary marginal troughs. Detection of a preceding Late Cretaceous inversion is complicated by the absence of sediments of this age. However, the Late Cretaceous deposits at the southern margin of the Weald contain gravity flows of allochthonous chalks and lateral changes in thickness and lithology, which are controlled by en-echelon folds of transpressional origin²⁴, indicating a Late Cretaceous uplift. The maximum uplift occurred where the thickest chalks had previously been deposited²⁴, yielding total erosion depths of approximately 1,500 m towards the east.

In the southern North Sea, several Palaeozoic and Mesozoic rift structures (the West Netherlands Basin^{3,19,25}, the Broad Fourteens^{3,19,26}, the Central Netherlands Basin³ and the Sole Pit Trough²⁷) experienced strong fault-controlled Late Cretaceous^{3,25} primary inversion resulting in uplift and 2,000–3,000-m-deep erosion of narrow ridges, and formation of associated primary marginal troughs (Fig. 3b). The onset of the Middle Palaeocene inversion phase is, as in the Danish Basin¹², marked by an extremely high percentage of reworked Late Cretaceous nannofossils in the Middle Palaeocene deposits²⁸. The distribution of the post-Danian Palaeocene sediments in the Netherlands and offshore reveals a broad, domal uplift and subsidence pattern distinctly different from the narrower, ridge-like structural style of the Late Cretaceous compressional inversion phase (Fig. 3b). Rather than additional compression, a Middle Palaeocene relaxation of in-plane tectonic stress excited the longer wavelength of vertical surface movements involved in flexural isostasy of the lithosphere, and averaged the loading effect of individual narrow inversion ridges into one broad flexural response. Continued regional subsidence in the area with differential loading of the crest and flanking marginal troughs by

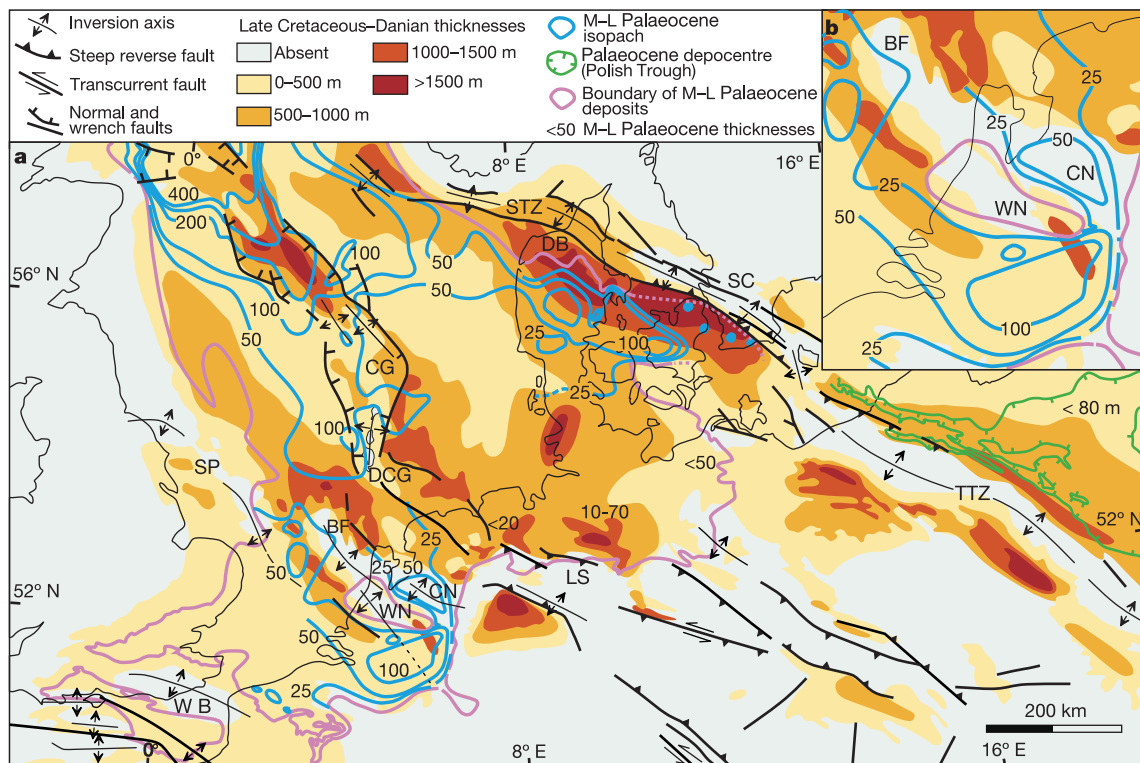


Figure 3 Isopach signature of inversion movements. **a**, Late Cretaceous–Danian isopach³ (primary inversions) with superimposed Middle–Late (M–L) Palaeocene depocentres³⁰ (secondary inversions) and inversion structures^{3,17–19}. BF, Broad Fourteens Basin; CN, Central Netherlands Basin; CG, Danish Central Graben; DB, Danish Basin; DCG, Dutch Central Graben; LS, Lower Saxony Basin; SC, Scania; SP, Sole Pit High;

STZ, Sorgenfrei-Tornquist Zone; TTZ, Tornquist-Teisseyre Zone; WB, Weald-Boulonnais area; WN, West Netherlands Basin. The northwestern end of the STZ appears asymmetric because of deep Quaternary erosion to the north. Around Scania the STZ inversion is asymmetric. The depocentre east of the TTZ represents Palaeocene deposits³¹. <50: M–L Palaeocene thickness less than 50 m. **b**, Enlarged view of southern North Sea.

the overlying sediment packages led to the apparent evolution of the domal inversion through the Eocene^{19,25}. The estimated 200–300 m of crestal erosion during this phase²⁵ agrees well with the predicted relative magnitudes of erosion for the primary and secondary inversion (Fig. 2). Relaxation of compressional induced stresses by ductile flow processes in the softer parts of the lithosphere may have contributed to the Palaeogene evolution^{9–11}. However, the plate-wide synchronicity of the Middle Palaeocene inversions indicates a sudden plate-wide release of inversion potential, a significant fraction of which must therefore have been maintained in compressional over-deepened elastic flexures.

An erosional phase in the southern North Sea around the Eocene–Oligocene boundary has been explained by a Pyrenean phase of compression^{3,19,25–27}, but may not be easy to distinguish from the erosional effects of the global sea level fall at that time²⁹.

It is widely accepted that flexural isostasy provides a causal relation between in-plane stress variations, differential vertical movements of the surface of the lithosphere and changes in the depositional patterns of the sedimentary sequences¹³. Application of this concept requires knowledge about the initial lithospheric deflection, which generally is not available away from topographic loads in the interior of continents. The central regions of rift basins, however, tend to become positively loaded during the rifting and thermal sag phase and therefore develop an intrinsic downward flexure, which in some cases has been revealed by accelerated subsidence¹⁵ in the incipient stage of compressional inversion. This pre-loading of rift basins promotes the formation of the asymmetric, primary marginal troughs during compressional inversion, and contributes positively to the over-deepening of the flexure that drives the relaxation inversion.

The existence of a positive load of the lithosphere along inversion ridges allows for prediction of the flexural response of the lithosphere to in-plane stress variations. Application of this concept to key European inversion structures allows us to explain the Middle Palaeocene change of inversion style, the synchronicity of the event across the European plate, the relative magnitude of the Late Cretaceous and Palaeocene erosion phases (when estimates exist) and the control exerted by the inversion movements on the post-Danian Palaeocene deposits.

The surprising conclusion that relaxation of in-plane tectonic stress, rather than increased Alpine compression, explains the European Middle Palaeocene inversions calls for reconsideration of the sources of the stresses affecting the interior of the European plate. Candidates for producing a plate-scale change of tectonic setting at this time involving forces in the range of 10^{12} N m⁻¹ include uplift of the North Atlantic lithosphere by the Iceland plume⁶ and the drop in north–south convergence rate between Africa and Europe⁷. □

Received 20 December 2004; accepted 23 March 2005; doi:10.1038/nature03599.

- Voigt, E. Über Randtröge vor Schollenrändern und ihre Bedeutung im Gebiet der Mitteleuropäischen Senke und angrenzender Gebiete. *Z. Deutsch. Geol. Gesell.* **114**, 378–418 (1962).
- Ziegler, P. A. Late Cretaceous and Cenozoic intra-plate compressional deformations in the Alpine foreland – a geodynamic model. *Tectonophysics* **137**, 389–420 (1987).
- Ziegler, P. A. *Geological Atlas of Western and Central Europe* 2nd edn (Shell International Petroleum Maatschappij B.V., Geological Society Publishing House, Bath, 1990).
- Ziegler, P. A., Cloetingh, S. & van Wees, J. D. Dynamics of intra-plate compressional deformation: the Alpine foreland and other examples. *Tectonophysics* **252**, 7–22 (1995).
- Ziegler, P. A., van Wees, J. D. & Cloetingh, S. Mechanical controls on collision-related compressional intraplate deformation. *Tectonophysics* **300**, 103–129 (1998).
- White, N. J. & Lovell, B. Measuring the pulse of a plume with the sedimentary record. *Nature* **387**, 888–891 (1997).
- Rosenbaum, G., Lister, G. S. & Duboz, C. Relative motions of Africa, Iberia and Europe during Alpine orogeny. *Tectonophysics* **359**, 117–129 (2002).
- Erlström, M., Thomas, S. A., Deeks, N. & Sived, U. Structure and tectonic evolution of the Tornquist Zone and adjacent sedimentary basins in Scania and the southern Baltic Sea area. *Tectonophysics* **271**, 191–225 (1997).
- Nielsen, S. B. & Hansen, D. L. Physical explanation of the formation and evolution of inversion zones and marginal troughs. *Geology* **28**, 875–878 (2000).
- Nielsen, D. L., Nielsen, S. B. & Lykke-Andersen, H. The post-Triassic evolution of the Sorgenfrei-Tornquist Zone – results from thermo mechanical modelling. *Tectonophysics* **328**, 245–267 (2000).
- Hansen, D. L. & Nielsen, S. B. Why rifts invert in compression. *Tectonophysics* **373**, 5–24 (2003).
- Clemmensen, A. & Thomsen, E. Palaeoenvironmental changes across the Danian-Selandian boundary

- in the North Sea Basin. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **219**, 351–394 (2005).
- Lambeck, K., Cloetingh, S. & McQueen, H. Intraplate stress and apparent changes in sea level: the basins of Northwestern Europe. *Can. Soc. Petrol. Geol. Mem.* **12**, 259–268 (1987).
- Gregersen, S., Voss, P. & TORWorking Group. Summary of project TOR: delineation of a stepwise, sharp, deep lithosphere transition across Germany–Denmark–Sweden. *Tectonophysics* **360**, 61–73 (2002).
- Dadlez, R., Narkiewicz, M., Stephenson, R. A., Visser, M. T. M. & Van Wees, D.-J. Tectonic evolution of the Mid-Polish Trough: modelling implications and significance for central European geology. *Tectonophysics* **252**, 179–195 (1995).
- Lamarche, J., Scheck, M. & Lewerenz, B. Heterogeneous inversion of the Mid-Polish Trough related to crustal architecture, sedimentary patterns and structural inheritance. *Tectonophysics* **373**, 141–159 (2003).
- Vejbæk, O. V. & Andersen, C. Post mid-Cretaceous inversion tectonics in the Danish Central Graben. *Bull. Geol. Soc. Denmark* **49**, 129–144 (2002).
- Clausen, O. R. & Korstgård, J. A. Tertiary tectonic evolution along the Arne-Elin Trend in the Danish Central Trough. *Terra Nova* **5**, 233–243 (1993).
- de Jager, J. Inverted basins in the Netherlands, similarities and differences. *Neth. J. Geosci.* **4**, 355–366 (2003).
- Baldschuhn, R., Best, G. & Kockel, F. N. in *Generation, Accumulation and Production of Europe's Hydrocarbons* (ed. Spencer, A. M.) 149–159 (Spec. Publ. 1, European Association of Petroleum Geoscientists, Oxford Univ. Press, Oxford, 1991).
- Kockel, F. Inversion structures in Central Europe—Expressions and reasons, an open discussion. *Neth. J. Geosci.* **82**, 367–382 (2003).
- Lake, S. D. & Karner, G. G. The structure and evolution of the Wessex Basin, southern England: an example of inversion tectonics. *Tectonophysics* **137**, 347–378 (1987).
- Jolley, D. W. Palynostratigraphy and depositional history of the Paleocene Ormesby/Thanet depositional sequence set in southeastern England and its correlation with continental West Europe and the Lista Formation, North Sea. *Rev. Palaeobot. Palynol.* **99**, 265–315 (1998).
- Mortimore, R. N. & Pomerol, B. Upper Cretaceous tectonic phases and end Cretaceous inversion in the Chalk of the Anglo-Paris Basin. *Proc. Geol. Ass.* **108**, 231–255 (1997).
- Worum, G. & Michon, L. Implications of continuous structural inversion in the West Netherlands Basin for understanding controls on Palaeogene deformation in NW Europe. *J. Geol. Soc. Lond.* **162**, 73–85 (2005).
- de Lugt, I. R., van Wees, J. D. & Wong, Th. E. in *Dynamics of Sedimentary Basin Inversion: Observations and Modelling* (eds Nielsen, S. B. & Bayer, U.) *Tectonophysics* **373** (Spec. Iss.), 141–159 (2003).
- van Hooft, B. Structural evolution, timing and tectonic style of the Sole Pit inversion. *Tectonophysics* **137**, 239–284 (1987).
- Steurbaert, E. High-resolution chronostratigraphy of Middle Paleocene to Early Eocene strata in Belgium and adjacent areas. *Palaeontographica Abt. A* **247**, 91–156 (1998).
- Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
- Vinken, R. et al. The Northwest European Tertiary Basin. Isopach map of Middle-Late Paleocene deposits. *Geol. Jb.* **100**, 1–508 (1988).
- Dadlez, R., Marek, S. & Pokorski, J. *Geological Map of Poland Without Cainozoic Deposits* (Polish Geological Institute, Warsaw, 2000).

Acknowledgements This work was completed during project CENMOVE, financed by the Danish Natural Science Research Council. D.L.H. thanks the Carlsberg Foundation for support. L. Mackay and C. Beaumont are thanked for reviews. M. Jarosinski, F. Kockel, J. Korstgaard, R. Mortimore, V. Otto, P. Poprawa and O. Vejbæk are thanked for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.B.N. (sbn@geo.au.dk).

Palaeomagnetism of the Vredefort meteorite crater and implications for craters on Mars

Laurent Carporzen¹, Stuart A. Gilder¹ & Rodger J. Hart²

¹*Institut de Physique du Globe, Laboratoire de Paléomagnétisme, 75252 Paris Cedex 05, France*

²*University of the Witwatersrand, Schonland Research Center, Wits 2050, Johannesburg, South Africa*

Magnetic surveys of the martian surface have revealed significantly lower magnetic field intensities over the gigantic impact craters Hellas and Argyre than over surrounding regions¹. The reduced fields are commonly attributed to pressure demagnetization caused by shock waves generated during meteorite impact^{2,3}, in the absence of a significant ambient magnetic

field. Lower than average magnetic field intensities are also observed above the Vredefort meteorite crater in South Africa, yet here we show that the rocks in this crater possess much higher magnetic intensities than equivalent lithologies found elsewhere on Earth. We find that palaeomagnetic directions of these strongly magnetized rocks are randomly oriented, with vector directions changing over centimetre length scales. Moreover, the magnetite grains contributing to the magnetic remanence crystallized during impact, which directly relates the randomization and intensification to the impact event. The strong and randomly oriented magnetization vectors effectively cancel out when summed over the whole crater. Seen from high altitudes, as for martian craters, the magnetic field appears much lower than that of neighbouring terranes, implying that magnetic anomalies of meteorite craters cannot be used as evidence for the absence of the planet's internally generated magnetic field at the time of impact.

The two-billion-year-old⁴ Vredefort structure (Fig. 1) represents the vestiges of an immense (>300 km original diameter) impact crater, where rebound and subsequent erosion have unearthed 1,000 km² of well-exposed Archaean crystalline basement. As the Earth's largest and oldest known impact structure, Vredefort serves as the best terrestrial analogue for the large martian craters. Shock deformation, in the form of planar fractures and planar deformation features (PDFs) in quartz, is abundant throughout the Vredefort rocks, and provides evidence for shock pressures exceeding 30 GPa (refs 5, 6). Pseudotachylite breccias and impact melt dykes are ubiquitous throughout the crater. The shocked rocks contain original Archaean multidomain magnetite grains, as well as single-domain-sized magnetite particles crystallized within the PDFs⁷. High magnetic intensities in shocked granite-gneisses were attributed to the single-domain magnetite^{8,9}. Paradoxically, the Earth's magnetic field is reduced by up to 4,000 nT at 300 m altitude over the area containing the highest magnetic intensities, which we will show is incompatible with shock demagnetization as suggested for the martian craters^{2,3}. To explain this contradiction and the origin of the high remanence in these rocks, we carried out a palaeomagnetic study of the Vredefort crater.

We drilled three or more palaeomagnetic cores at 127 sites throughout the crater in the granite-gneisses, as well as in the pseudotachylites and impact dykes that cut them (Fig. 1). Each core was oriented with Sun and magnetic compasses. Strongly magnetized crust made Sun compass corrections necessary, as 22% of the measurements deviate more than 10° from the international geomagnetic reference field. Thermomagnetic experiments characterize Curie points of $577 \pm 3^\circ\text{C}$ ($N = 3$) in the granite-gneisses and $571 \pm 1^\circ\text{C}$ ($N = 4$) in the granulites, which correspond to Fe-pure magnetite and magnetite containing ~2% Ti, respectively.

One granite-gneiss sample, cut from one core per site, was subjected to stepwise alternating field (AF) demagnetization that normally isolated a very well defined component (Fig. 2a and b), whose direction was randomly oriented from one site to the next (Fig. 3a). Following these results, we thermally demagnetized samples from 31 cores, again isolating well defined linear magnetization components (Fig. 2c), with randomly distributed directions (Fig. 3b). The average angular distance between the magnetic vectors of two samples cut from two different cores from the same site is $64.7 \pm 45.9^\circ$. As the average distance between two cores from the same site is 30 cm, the origin of the dispersion is less than a few decimetres.

To better define the length scale of the dispersion, we compared the magnetization directions between two samples cut from the same core for 34 cores (68 AF demagnetized samples). The average angular distance between magnetic vectors of the two samples from the same core is $9.4 \pm 6.8^\circ$ (omitting two outliers). This value remains high with respect to rocks possessing thermal remanent

magnetizations (TRMs), which are rarely different by a degree, and suggests that the length scale of the dispersion lies within the dimension of a specimen (2.2 cm).

We then measured the anisotropy of magnetic susceptibility (AMS) on the same samples that underwent demagnetization. Interestingly, like the magnetic directions, principal susceptibility axes are randomly oriented on the site level (Fig. 3e), and the dispersion decreases as the length scale diminishes (Fig. 4). This is an important finding, as the magnetic foliations do not correlate with observable magmatic and metamorphic foliations, which trend consistently over distances greater than several metres. Because each sample contains multiple randomly oriented quartz grains that in turn contain multiple PDF geometries, the overall magnetic fabric is the sum of several individual fabrics that overprint the original magnetic foliation. If the AMS is controlled by single-domain magnetite, then the anisotropy of isothermal remanent magnetization (aIRM) should be defined by the same grains. Of nine aIRM experiments, the AMS and aIRM principal-axis directions are the same in seven samples whose anisotropies are high (9–29%) and similar in two cases with lower anisotropies (5%), meaning that the magnetite grains controlling the AMS also carry the magnetic remanence. Whereas most single-domain magnetite-bearing terrestrial rocks have inverse or intermediate AMS fabrics, Vredefort does not, owing to multiple individual planes of single-domain magnetite in the PDFs. Thus, the single-domain magnetite grains created during impact control the AMS and carry the magnetic remanence. Although lightning could potentially deviate the magnetization directions, it cannot reorient the magnetic anisotropy of the granite-gneisses. We conclude that the dispersion of the magnetization and AMS directions in the granite-gneisses is a characteristic feature of the impact event.

The magnetic behaviour is different for the melt rocks. A coherent direction is isolated at higher unblocking coercivities in 30 of 43 samples from five dykes (12 sites) distributed throughout the crater (declination $D = 25.0^\circ$, inclination $I = 57.2^\circ$, $\alpha_{95} = 3.9^\circ$ (where α_{95} is the radius that the mean direction lies within 95% confidence); Figs 1, 2d and 3c). The remaining 13 samples, which come from the same dyke, have higher natural remanent magnetization (NRM) intensities ($20.8 \pm 5.8 \text{ A m}^{-1}$) than those possessing coherent directions ($6.2 \pm 5.7 \text{ A m}^{-1}$). The samples with high NRM intensities have possibly been hit by lightning. Nineteen of 23 pseudotachylite samples display a linear magnetization component (Fig. 2e), whose directions are well clustered within and

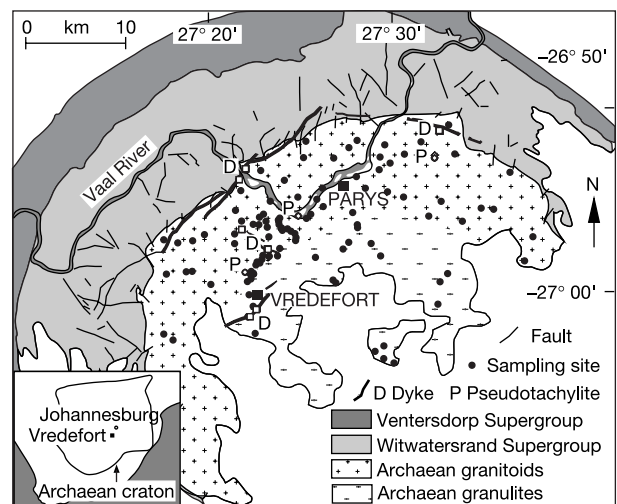


Figure 1 Simplified geologic map of the Vredefort crater (after ref. 25).

between sites, defining a mean of $D = 20.8^\circ$, $I = 57.4^\circ$ ($\alpha_{95} = 1.9^\circ$, Fig. 3d). The four pseudotachylite samples whose directions lie far from the dominant population have higher NRM intensities ($16.9 \pm 10.8 \text{ A m}^{-1}$) than the rest ($1.8 \pm 1.4 \text{ A m}^{-1}$). Again, we suspect that lightning is the cause for the aberrant directions in these latter samples.

Other palaeomagnetic studies in the Vredefort crater have also found consistent directions in the impact melt rocks^{10,11}, but basement rocks with high angular dispersions were rejected in the final analyses. Besides Vredefort, palaeomagnetic studies of other impact craters have found wildly variable magnetization directions in shocked facies^{12,13}. Until now, no explanation has been given for the origin of these aberrant directions.

The shocked samples have extremely high NRM intensities and Koenigsberger (Q)-ratios (here defined as $[\text{NRM}/(\text{susceptibility} \times 23.9 \text{ A m}^{-1})]$) compared with similar rock types found worldwide¹⁴. NRMs range from 0.01 to 110.0 A m^{-1} (average, 16.2 A m^{-1}) and Q-ratios range from 0.2 to 748.7 (average, 107.7) for 182 samples. For comparison, NRM intensities and Q-ratios for

the Cretaceous Mt Givens granite¹⁵ (Sierra Nevada, California, $N = 247$) range from 0.01 to 7.3 A m^{-1} (average, 0.4 A m^{-1}) and from 0.02 to 28.9 (average, 1.1); mean NRM and Q-ratios for the Archaean Pomovaara granite¹⁶ (northern Finland) are 0.2 A m^{-1} and 0.5 ($N = 141$). Because 26% of the dyke and pseudotachylite samples were probably struck by lightning, the same proportion of granites could have also been affected. Subtracting 26% of the samples with the highest intensities yields a range in NRMs from 0.01 to 22.0 A m^{-1} (average, 5.7 A m^{-1}) and Q-ratios from 0.2 to 366.6 (average, 71.2) for $N = 135$; this subpopulation also has randomly oriented magnetic directions. We thus concur with Hart *et al.*⁸ that the NRM intensities and Q-ratios of the Vredefort basement rocks are elevated compared to similar rock types found elsewhere.

Our study shows that impact acts to randomize and increase the magnetization in shocked rocks. The main difference between the shocked and non-shocked rocks is the relative timing of the lock-in of magnetic remanence. Pseudotachylites and dykes at Vredefort possess TRMs with magnetic characteristics similar to those of volcanic rocks that cooled in the Earth's magnetic field. Magnetic microscopy of the shocked facies identified preferential alignment of the magnetic moments within PDF planes⁷, showing that there was an ambient magnetic field present when the magnetite grains acquired their remanences. However, because there is no hemispherical bias in the recorded palaeomagnetic directions (Fig. 3a and b), the external field direction could not have been constant over the crater. Thus, although PDFs control the distribution of magnetite in shocked rocks, an external field, whose orientation varied over centimetre length scales, must have been responsible for aligning their magnetic remanences. This is consistent with laboratory experiments, which show that plasma produced from impact events generates intense magnetic fields¹⁷. A 1-km-diameter bolide has been calculated to produce a 0.03-T field (1,000 times greater than

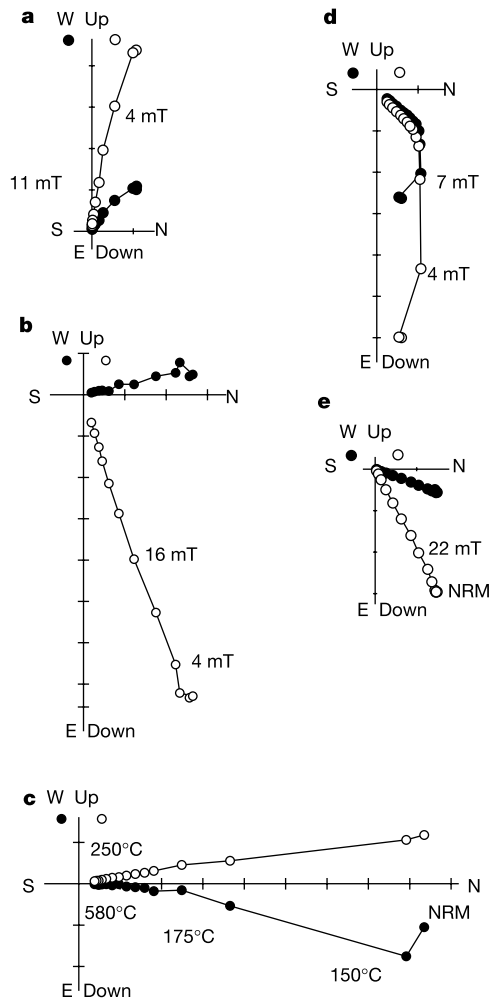


Figure 2 Representative demagnetization diagrams of Vredefort rocks. **a**, **b**, Alternating field (AF) demagnetization of shocked granite-gneiss; **a**, sample V4085A; **b**, V217A. **c**, Thermal demagnetization of shocked granite-gneiss (V047A). **d**, AF demagnetization of a dyke (V082B). **e**, AF demagnetization of a pseudotachylite (V4047A). Filled (open) symbols are the horizontal (vertical) magnetic component with 'up' and 'down' specifying inclination direction; N, S, E and W represent respectively north, south, east and west; NRM, natural remanent magnetization; specified demagnetization steps are in milliteslas (mT) or degrees Celsius ($^\circ\text{C}$).

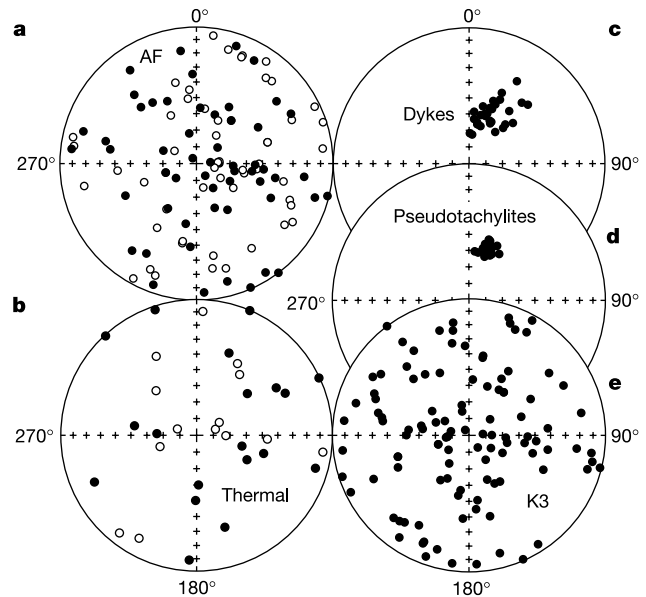


Figure 3 Equal area stereonet plots of the magnetic measurements. **a**, High coercivity magnetization component directions of 111 shocked granitoid samples. **b**, High temperature magnetization component of 31 shocked granitoid samples. **c**, High coercivity magnetization component of 30 dyke samples. **d**, High coercivity magnetization component of 23 pseudotachylite samples. **e**, Directions of minimum anisotropy of magnetic susceptibility axes (K3) of the same 111 shocked samples shown in **a**. Filled (open) circles are lower (upper) hemisphere projections of the magnetization component directions in **a** to **d**. In **e**, all minimum susceptibility axes directions are plotted on the lower hemisphere.

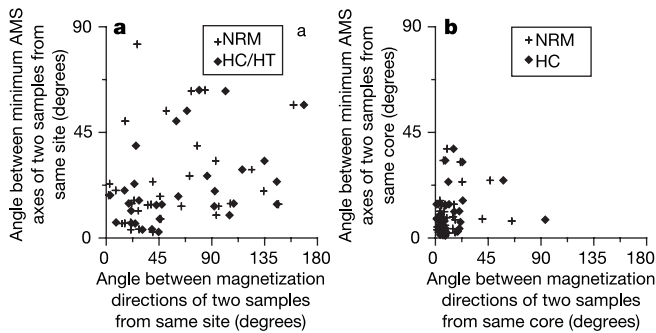


Figure 4 Comparison of the angular distance between minimum anisotropy of magnetic susceptibility (AMS) axes and magnetization component directions between two samples. **a**, Two samples from the same site; **b**, two samples from the same core. NRM, natural remanent magnetization direction; HC/HT, high coercivity or high temperature remanent magnetization direction; both are shown for comparison.

the Earth's) at a distance of 100 km that lasts for 100 s after an impact¹⁸. For Vredefort, these are minimum values because the meteorite diameter was greater than 10 km (ref. 19).

We suggest that small-wavelength magnetic fields of high intensity are responsible for both the randomization and the unusually high remanence of the Vredefort basement rocks. Diamond anvil experiments on magnetite show that its remanent saturation moment is significantly enhanced when it is either under pressure or after decompression²⁰. This implies that shock creates a piezo-remanent magnetization, or a high-field TRM, depending on how fast the newly formed magnetite grains cooled through the Curie temperature, which is 580 °C at 0.1 MPa yet 1,180 °C at 30 GPa (refs 21, 22). Both processes would increase, not decrease, the magnetic intensities of the shocked rocks. Assuming that the temperature of the pre-impact crust was 500 °C, magnetite grains crystallizing in a 1-μm-thick PDF plane would cool through the Curie point in less than one second^{7,23}, in the plasma generated field. In contrast, a 1-m-wide dyke takes about 11 days to cool²³, long after the plasma field disappeared, in the presence of a 'normal' Earth field.

Aeromagnetic surveys over meteorite impacts commonly detect field intensities lower than regional magnetic trends²⁴. Such is also the case for the gigantic (>1,000 km diameter) martian impact basins Hellas and Argyre¹, whose lower field intensities are commonly ascribed to pressure demagnetization^{2,3}. Our study of Vredefort, the best terrestrial analogue of a martian meteorite crater, demonstrates that shock in fact leads to intensified and randomized magnetic vectors in shocked rocks. On the scale of the crater, the vectors cancel out, being equivalent to a volume of non-magnetic material seen at high altitudes, thus accounting for the reduced fields measured over the craters. The lack of crustal magnetism over impact basins on Mars is commonly ascribed to the absence of an internal dynamo during meteorite impact. Our new findings question the validity of this assumption. □

Received 18 October 2004; accepted 10 March 2005; doi:10.1038/nature03560.

1. Acuña, M. H. *et al.* Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. *Science* **284**, 790–793 (1999).
2. Hood, L. L., Richmond, N. C., Pierazzo, E. & Rochette, P. Distribution of crustal magnetic fields on Mars: Shock effects of basin-forming impacts. *Geophys. Res. Lett.* **30**, 1281, doi:10.1029/2002GL016657 (2003).
3. Mohit, P. S. & Arkani-Hamed, J. Impact demagnetization of the Martian crust. *Icarus* **168**, 305–317 (2004).
4. Moser, D. E. Dating the shock wave and thermal imprint of the giant Vredefort impact, South Africa. *Geology* **25**, 7–10 (1997).
5. Stöffler, D. & Langenhorst, F. Shock metamorphism of quartz in nature and experiment: I. Basic observation and theory. *Meteoritics* **29**, 155–181 (1994).
6. Grieve, R. A. F., Langenhorst, F. & Stöffler, D. Shock metamorphism of quartz in nature and experiment: II. Significance in geoscience. *Meteoritics* **31**, 6–35 (1996).

7. Cloete, M. *et al.* Characterization of magnetite particles in shocked quartz by means of electron- and magnetic force microscopy: Vredefort, South Africa. *Contrib. Mineral. Petrol.* **137**, 232–245 (1999).
8. Hart, R. J., Hargraves, R. B., Andreoli, M. A. G., Tredoux, M. & Doucourné, C. M. Magnetic anomaly near the center of the Vredefort structure: implications for impact-related magnetic signatures. *Geology* **23**, 277–280 (1995).
9. Hart, R. J. *et al.* 'Super magnetic' rocks generated by shock metamorphism from the center of the Vredefort impact structure, South Africa. *S. Afr. J. Geol.* **103**, 151–155 (2000).
10. Hargraves, R. B. Paleomagnetic evidence relevant to the origin of the Vredefort ring. *J. Geol.* **78**, 253–263 (1970).
11. Jackson, G. M. *Paleomagnetic Study and Magnetic Modeling of the Vredefort Dome*. Thesis, Univ. Witwatersrand (1982).
12. Hargraves, R. B. & Roy, D. W. Paleomagnetism of anorthosite in and around the Charlevoix cryptoexplosion structure, Quebec. *Can. J. Earth Sci.* **11**, 854–859 (1974).
13. Halls, H. C. Shock-induced remanent magnetization in late Precambrian rocks from Lake Superior. *Nature* **255**, 692–695 (1975).
14. Clark, D. A. & Emerson, D. W. Notes on rock magnetic characteristics in applied geophysical studies. *Exp. Geophys.* **22**, 547–555 (1991).
15. Gilder, S. & McNulty, B. Tectonic exhumation and tilting of the Mount Givens pluton, central Sierra Nevada, California. *Geology* **27**, 919–922 (1999).
16. Wennerström, M. & Airo, M. L. Magnetic fabric and emplacement of the post-collisional Pomavaara Granite Complex in northern Fennoscandia. *Lithos* **45**, 131–145 (1998).
17. Crawford, D. A. & Schultz, P. H. Laboratory observations of impact-generated magnetic fields. *Nature* **336**, 50–52 (1988).
18. Crawford, D. A. & Schultz, P. H. Electromagnetic properties of impact-generated plasma, vapor and debris. *Int. J. Impact Eng.* **23**, 169–180 (1999).
19. Turtle, E. P. & Pierazzo, E. Constraints on the size of the Vredefort impact crater from numerical modeling. *Meteorit. Planet. Sci.* **33**, 483–490 (1998).
20. Gilder, S. A., LeGoff, M., Chervin, J. C. & Peyronneau, J. Magnetic properties of single and multi-domain magnetite under pressures from 0 to 6 GPa. *Geophys. Res. Lett.* **31**, L10612, doi:10.1029/2004GL019844 (2004).
21. Samara, G. & Giardini, A. Effect of pressure on the Néel temperature of magnetite. *Phys. Rev.* **186**, 577–580 (1969).
22. Schult, A. Effect of pressure on the Curie temperature of titanomagnetites [(1 - x)Fe₃O₄-xTiFe₂O₄]. *Earth Planet. Sci. Lett.* **10**, 81–86 (1970).
23. Turcott, D. L. & Schubert, G. *Geodynamics* (Wiley & Sons, New York, 1982).
24. Grieve, R. A. F. & Pilkington, M. The signature of terrestrial impacts. *J. Austr. Geol. Geophys.* **16**, 399–420 (1996).
25. Nel, L. T. *Geological Map of the Country Around Vredefort. Scale 1:63360* (Union of S. Afr., Dept Mines and Ind., Geol. Surv., Pretoria, 1927).

Acknowledgements We thank A. Galdeano, G. Hulot, M. Cloete, M. Le Goff, Y. Gallet and G. Plenier for discussions and comments. This work was supported by the French Ministries of Foreign Affairs and Education and Research, INSU and the South African National Research Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.A.G. (gilder@ipgp.jussieu.fr).

Advanced optics in a jellyfish eye

Dan-E. Nilsson¹, Lars Gislén², Melissa M. Coates¹, Charlotta Skogh¹ & Anders Garm¹

¹Department of Cell and Organism Biology, Lund University, Zoology Building, Helgonavägen 3, 22362 Lund, Sweden

²Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE - 223 62 Lund, Sweden

Cubozoans, or box jellyfish, differ from all other cnidarians by an active fish-like behaviour and an elaborate sensory apparatus^{1,2}. Each of the four sides of the animal carries a conspicuous sensory club (the rhopalium), which has evolved into a bizarre cluster of different eyes³. Two of the eyes on each rhopalium have long been known to resemble eyes of higher animals, but the function and performance of these eyes have remained unknown⁴. Here we show that box-jellyfish lenses contain a finely tuned refractive index gradient producing nearly aberration-free imaging. This demonstrates that even simple animals have been able to evolve the sophisticated visual optics previously known only from a few

advanced bilaterian phyla. However, the position of the retina does not coincide with the sharp image, leading to very wide and complex receptive fields in individual photoreceptors. We argue that this may be useful in eyes serving a single visual task. The findings indicate that tailoring of complex receptive fields might have been one of the original driving forces in the evolution of animal lenses.

In the light of the current interest in early eye evolution^{5,6}, the uniquely evolved eyes of box jellyfish have been neglected. To a much higher degree than other jellyfish, cubozoan medusae actively position themselves in suitable habitats^{1,7}. This is the likely reason that they have eyes rivaling those of advanced Bilateria. Suspended from the sides of the bell, each box jellyfish carries four sensory clubs (rhopalia), and each of these is equipped with two lens eyes, here termed the upper and lower eyes, together with two different pairs of simpler pigment-pit eyes^{2,3,8–10} (Fig. 1), making a total of 24 eyes per animal. From the unique crystallin proteins we know that at least the lenses have evolved independently in box jellyfish^{11,12}. Making good lenses seems to be a demanding task, because only few animal phyla have accomplished it¹³. Our question here is whether the eyes of box jellyfish are simple sensory organs or whether they have, to any degree, managed to evolve the sophisticated optics found in vertebrate and cephalopod eyes¹⁴, giving them image-forming capabilities.

The eyes do not vary much between box jellyfish species, and for convenience we used a small (10 mm) Caribbean species, *Tripedalia cystophora*, which can be bred from polyps under laboratory conditions¹⁵. The natural habitat of this species is among the prop roots around the edge of mangrove lagoons¹. To investigate the visual optics we first established accurate geometrical models of the upper and lower lens eyes (Fig. 1d). These have diameters of 150 and 250 μm respectively. Both eyes are built similarly but differ somewhat in retinal geometry. All major components of a typical camera-type eye are present: a cornea, a lens, a retina, a pigment layer and an iris. The lenses are spherical and cellular. A thin cellular space separates the lens from the retina. Microvilli projecting from the ciliary outer segments of photoreceptor cells fill the retina³. The outer segments are contiguous and lack structures that can act as light guides. They are arranged in an orderly way but have their long axes unconventionally aligned towards a point just outside the pupil in the lower eye and obliquely towards the pupil margin in the upper eye (Fig. 1d). A layer of dark pigment covers the outside of the retina and forms an iris around the lens. Changing the light intensity reveals that the lower eye has a mobile pupil that can close down the aperture from about 150 μm to 100 μm in less than 1 min (Fig. 1), whereas the pupil of the upper eye remains constant at all intensities.

Fresh isolated lenses are able to form good images (Fig. 2) irrespective of lens orientation, implying spherically symmetrical optics. Using microinterferometry we determined the refractive index from the periphery to the centre of the lenses. From the data (Fig. 2) it is clear that the lenses of both eyes contain refractive index gradients. The central refractive index is about 1.48, which is a little lower than the 1.49–1.56 found in fish lenses^{14,16}. In the upper eye the lenses display a smooth gradual decrease in refractive index towards the periphery, whereas lenses of the lower eye have a nearly homogeneous core surrounded by a smooth gradient. In the lens periphery of both eyes the refractive index gradient falls without discontinuity to the value of the surrounding tissue (1.34). Electron microscopic sections of the cellular lenses reveal a corresponding staining gradient, indicating that the packing density of crystallin proteins is responsible for the gradient in refractive index.

Tracing rays through the refractive-index gradient of the upper eye reveal nearly perfect focusing for all ray positions (Fig. 2). For such a minute eye it is surprising to find well-corrected, aberration-free imaging, otherwise known only from the much larger eyes of

vertebrates and cephalopods^{13,14,16}. The gradient in the upper-eye lenses comes very close to the ideal solution¹⁷. The lenses of the lower eye have a less ideal gradient and consequently display some spherical aberration (Fig. 2e, f). It is the homogeneous lens core and steep peripheral gradient that results in positive spherical aberration in the lower eye.

When the ray paths of the lenses are put into the geometrical models of the eyes, it turns out that both eyes are severely under-focused (Fig. 3a). The sharp image falls well below the retina and it would seem that the sharp focus of the lenses is wasted by inappropriate eye geometry. Another, more likely, interpretation is that the eyes are 'purposely' under-focused to remove high spatial frequencies (fine image details) from the retinal image, much as occurs in insect dorsal ocelli¹⁸. If the arrangement is indeed a spatial

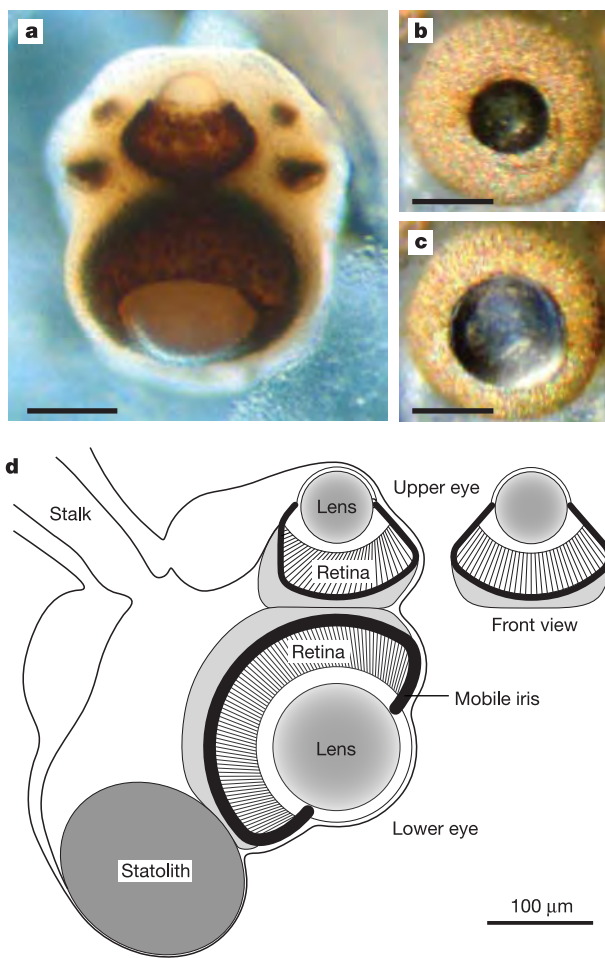


Figure 1 The eyes of the box jellyfish *Tripedalia cystophora*. **a**, The rhopalium shows the upper and lower lens eyes flanked by two pairs of simpler eyes. **b**, **c**, The live lower eye displays a mobile pupil. In **b** the eye was exposed for about 10 min to light intensities corresponding to direct sunlight, which is enough to close the pupil maximally. The fully open pupil in **c** is the result of total darkness for 10 min. Pupil adjustments take about 1 min. **d**, An accurate anatomical model. The sagittal section contains the statolith and the internal structure of the two lens eyes. The spherical lenses are surrounded by a cellular capsule, the inner part of which forms the equivalent of a vitreous body between lens and retina. Iris constriction in the large eye is caused by contraction of the outer part of the lens capsule. The lower eye is rotationally symmetrical, but the upper eye is only bilaterally symmetrical (front view shown to the right). Receptor outer segments fill the retina of both lens eyes. The alignment of receptor outer segments is unusual, especially in the upper eye, where receptor axes converge on a point at one side of the lens. Scale bars, 100 μm .

low-pass filter, it would help the animals to detect the large and stationary structures of their visual environment, but would leave unseen the plankton and small particles floating with the current. Assuming that the lens eyes have evolved to allow the jellyfish to remain in nearshore habitats and to avoid swimming into obstacles, a low-pass filtering of image structure would make sense.

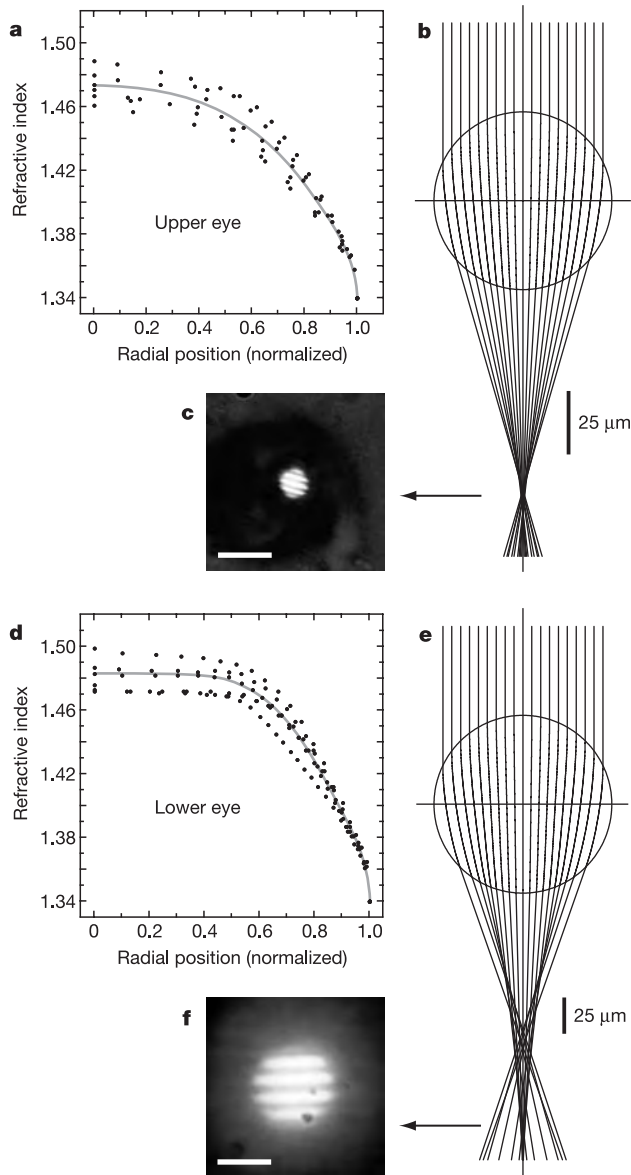


Figure 2 Optical properties of the lenses. **a–c**, Upper eye; **d–f**, lower eye. In **a** and **d** the refractive index is plotted as a function of radial position in the lenses (values from six lenses of each kind, normalized to unity radius). In the upper-eye lens (**a**) the refractive index follows a near-parabolic profile, whereas the lower-eye lens (**d**) has a homogeneous centre and a gradient only in the outer half. Curves fitted to the data (grey traces in **a** and **d**) were used to calculate the ray paths in **b** and **e**. From these it is obvious that the lens of the upper eye forms a much better image than that of the lower eye. The upper-eye lens produces an almost aberration-free focus at a distance of 3.3 radii from the lens centre, whereas the lower-eye lens displays positive spherical aberration, with focal lengths ranging from 2.6 to 3.7 lens radii. **c**, **f**, The difference in image quality was confirmed by direct observation of images produced behind fresh isolated lenses. A grating object at infinity generates a crisp image behind a lens of the upper eye (**c**), but the same object imaged by a lens from a lower eye is contaminated by considerable aberration blur (**f**). The difference in image magnification is a consequence of the different focal lengths. Scale bars, 25 μm.

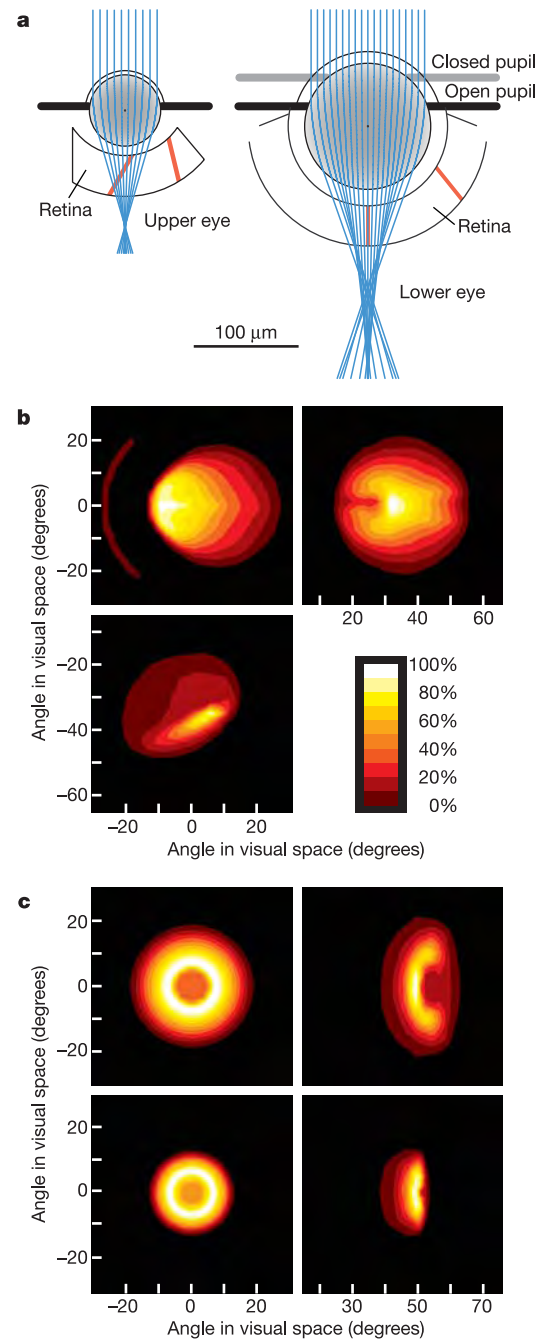


Figure 3 Modelling of receptive fields of individual photoreceptors in the retina of the upper and lower eyes. **a**, The optical models were used for tracing rays through the lens and retina and computing the absorption of light in selected single photoreceptors (red bars). Rays were traced in three dimensions, and by calculating receptor absorption at different incident angles of the ray bundle it was possible to generate receptive field maps for any receptor in the eyes. **b**, **c**, Receptive fields of receptors in the upper eye (**b**) and the lower eye (**c**). The top panels in **b** are for the central and peripheral receptors in the upper eye, as indicated in **a**. Note that the tilted receptors cause asymmetric receptive fields. The bottom panel in **b** shows the receptive field of a peripheral receptor in the perpendicular plane of the upper eye (see frontal view in Fig. 1d). Because the lower eye (**c**) is rotationally symmetrical, the perpendicular plane is identical with that shown in **a**. The top panels are modelled with an open pupil, for central (left) and peripheral (right) receptors (see red bars in **a**). The bottom panels are for the same receptors under a closed pupil. The sensitivity of the receptors was normalized to 100% in accordance with the colour map in **b**.

To analyse the consequence of under-focused optics and lack of light-guiding structures in the retina further, it is necessary to measure the angular sensitivity functions of individual photoreceptors. Normally, intracellular electrophysiological recordings would be the method of choice here, but we have not yet succeeded with this approach. However, our optical and anatomical data allow an alternative approach: calculations of light absorption in single receptors using a three-dimensional ray-path model. This approach has the advantage of allowing a direct correlation between the angular acceptance function of a receptor and its position in the retina. We traced three-dimensional beams of light in a fine angular matrix covering the visual field of the eye, and calculated the amount of light absorbed in selected photoreceptor cells as a function of the angle in visual space. The resulting maps are the complete angular sensitivity functions (receptive fields) of selected receptors (Fig. 3).

As expected from the severe under-focusing, the receptive fields are very wide compared with those of photoreceptors in other eyes. Half-widths of the receptive fields are about 20° , which is to be compared with a few degrees in insect compound-eye photoreceptors, and much less than 1° in vertebrate camera-eye photoreceptors¹³. If they had light-guiding photoreceptors in the focal plane of the lens, both the upper and lower eye of *Tripedalia* would easily be able to obtain angular sensitivities less than 1° . The visual resolution is thus very much worse than the sophisticated lenses would allow. Another peculiarity of the receptive fields is their shape. Central receptors in the lower eye display a ring of highest sensitivity surrounding a pronounced dip, and other receptors have strongly non-concentric sensitivity functions (Fig. 3). A recurring feature in our modelled receptive fields is that the sensitivity peak is flanked by a steep slope on one side and a much slower decrease on the other. This indicates that higher spatial frequencies at certain orientations remain in the otherwise very blurred image.

Because we do not yet know how the visual information from the lens eyes of box jellyfish is processed and used, we cannot tell what purpose the peculiar sensitivity functions might serve. But it is intriguing to note that many neurons in higher visual centres of the vertebrate brain also have large and geometrically complex receptive fields^{19–21}. A typical feature of animal visual systems is that higher processing occurs in parallel pathways where each pathway handles a specific aspect of information such as large-field motion detection or feature recognition²². The large and complex receptive fields of neurons found in vertebrate higher visual centres represent highly filtered information needed for specific visual tasks. In box jellyfish we find these large complex receptive fields at the level of photoreceptors, indicating that the eyes might be specialized for a specific task only and that this allows complex filtering of information much earlier than in more general visual systems. The fact that box jellyfish have four different types of eye gives support to the idea that each eye type is highly specialized.

The early evolution of animal visual systems is likely to have started out with eyes that were involved only in single visual tasks. In this perspective it is interesting to note that high visual acuity is not necessarily desirable. The lens eyes of box jellyfish indicate that there might be visual tasks best served by a blurred image. Evolution of sophisticated eyes might therefore be a process with discrete stages representing the sequential addition of visual tasks. Our results also indicate that advanced lenses with graded-index optics might have evolved for tailoring complex receptive fields and not just for improving sensitivity or acuity. □

Methods

Animals

Adult *Tripedalia cystophora* were collected in mangrove swamps near La Parguera, Puerto Rico. Experiments were performed on animals caught in the wild and on descendants cultured in the laboratory. We used medusae with a bell height of 6–8 mm.

Anatomical model

To obtain an anatomically precise model of the eye geometry we used sections of rhopalia prepared in fixative (2.5% paraformaldehyde, 2.5% glutaraldehyde and 3% sucrose in 0.1 M cacodylate buffer pH 7.4) for 2 h, dehydrated in an ethanol series and embedded in Epon resin. The plastic blocks were cut at $0.5\ \mu\text{m}$ on a diamond knife and stained with methylene blue and azur blue. Sections for light microscopy were cut through the sagittal plane of the rhopalium and perpendicular to this plane along the optical axis of the upper and lower lens eyes respectively. Geometrical models were generated as an average based on eight rhopalia from different animals. The sections suffered some shrinkage and minor deformations, which was corrected in the model by comparison with photographs of 12 fresh rhopalia in which the outer margins of the lens and retina were visible. Photoreceptor dimensions were measured on material prepared as above but block-stained with OsO_4 , sectioned for electron microscopy, post-stained with lead citrate and uranyl acetate, and examined in a Jeol transmission electron microscope.

Micointerferometry

Fresh lenses were dissected out of rhopalia with thin tungsten needles. The success rate was low, and only about one in every five trials resulted in an intact isolated lens. Dissections and subsequent measurements were made in marine salt water. The isolated lenses were placed in a parallel-sided cavity of strain-free glass, wide enough not to squeeze the lens. The cavity was placed in a Zeiss interference microscope (Jamin-Lebedeff type) fitted with interference condenser and objective for $\times 10$ magnification. The measurement beam passed through the jellyfish lens, and the reference beam passed immediately next to it. Monochromatic light of wavelength 546 nm was used. The microscope was focusing at the widest part of the jellyfish lens, and the background phase shift was adjusted to maximum extinction. This double-beam interferometric setup produced concentric interference fringes in the lenses. From the position of the fringes, phase shift profiles were generated in 5 to 12 different directions from the centre to the periphery of each lens. Average phase shift profiles for each lens were then converted to refractive index profiles. The conversion, which assumes rotational symmetry, was performed with an iterative ray-tracing procedure²³. Refractive index profiles were calculated for lenses of six upper and six lower eyes, and average profiles were fitted by cubic-spline functions.

Optical modelling

To generate receptive fields (Fig. 3), a matrix of 100,000 parallel rays, evenly spaced over $120\ \mu\text{m} \times 120\ \mu\text{m}$, were traced through the full pupil apertures of the upper and lower lens eye, using the measured refractive index gradients (Fig. 2) and the geometrical eye models (Figs 1d, 3a). The trajectory of each ray was calculated in steps of $0.05\ \mu\text{m}$, accounting for refraction at interfaces and in the lens gradient (see ref. 23 for equations). The water surrounding the eye was given a refractive index of $n = 1.33$, the cellular lens capsule $n = 1.34$ and the retinal receptors $n = 1.36$. The retina was assumed to absorb light with an absorption coefficient of $0.01\ \mu\text{m}^{-1}$, which is a typical value for receptive segments of photoreceptors with microvilli²⁴. Receptors were modelled as cylindrical structures, and absorption in a selected target receptor provided the calculated output from each ray tracing. The incident angle of the ray matrix was changed in increments of 1° in one plane, and the other plane was accessed by rotating the arc of different incident angles azimuthally in steps of 5° . Data on the receptive field are thus most densely sampled in the centre of the visual field. The ray-tracing program was written in Pascal (Metrowerks CodeWarrior), and the response data were plotted by using Mathworks MATLAB with cubic-spline interpolation.

Received 29 November 2004; accepted 21 February 2005; doi:10.1038/nature03484.

- Buskey, E. J. Behavioral adaptations of the cubozoan medusa *Tripedalia cystophora* for feeding on copepods (*Dioithona oculata*) swarms. *Mar. Biol.* **142**, 225–232 (2003).
- Berger, E. W. The histological structure of the eyes of cubomedusae. *J. Comp. Neurol.* **8**, 223–230 (1898).
- Laska, G. & Hündgen, M. Morphologie und Ultrastruktur der Lichtsinnesorgane von *Tripedalia cystophora* Conant (Cnidaria, Cubozoa). *Zool. Jb. Anat.* **108**, 107–123 (1982).
- Coates, M. M. Visual ecology and functional morphology of the Cubozoa. *Integr. Comp. Biol.* **43**, 542–548 (2003).
- Nilsson, D.-E. Eye evolution: a question of genetic promiscuity. *Curr. Opin. Neurobiol.* **14**, 407–414 (2004).
- Arendt, D., Tessmar-Raible, K., Snyman, H., Dorrestein, A. W. & Wittbrodt, J. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science* **306**, 869–871 (2004).
- Stewart, S. E. Field behavior of *Tripedalia cystophora* (Class Cubozoa). *Mar. Freshwater Behav. Physiol.* **27**, 175–188 (1996).
- Martin, V. J. Photoreceptors of cnidarians. *Can. J. Zool.* **80**, 1703–1722 (2002).
- Piatigorsky, J. & Kozmik, Z. Cubozoan jellyfish: an evo/devo model for eyes and other sensory systems. *Int. J. Dev. Biol.* **48**, 719–729 (2004).
- Laska, G. & Hündgen, M. Die Ultrastruktur des neuromuskulären Systems der Medusen von *Tripedalia cystophora* und *Carybdea marsupialis* (Coelenterata, Cubozoa). *Zoomorphology* **104**, 163–170 (1984).
- Piatigorsky, J., Horwitz, J. & Norman, B. L. J1-crystallins of the cubomedusan jellyfish lens constitute a novel family encoded in at least three intronless genes. *J. Biol. Chem.* **268**, 11894–11901 (1993).
- Piatigorsky, J. et al. J3-crystalline of the jellyfish lens: Similarity to saposins. *Proc. Natl Acad. Sci. USA* **98**, 12362–12367 (2001).
- Land, M. F. & Nilsson, D.-E. *Animal Eyes* (Oxford Univ. Press, Oxford, 2002).
- Kröger, R. H. H., Campbell, C. W., Munger, R. & Fernald, R. D. Refractive index distribution and spherical aberration in the crystalline lens of the African cichlid fish *Haplochromis burtoni*. *Vision Res.* **34**, 1815–1822 (1994).
- Werner, B., Cutress, C. E. & Studebaker, J. P. Life cycle of *Tripedalia cystophora* Conant (Cubomedusae). *Nature* **232**, 582–583 (1971).
- Jagger, W. S. & Sands, P. J. A wide-angle gradient index optical model of the crystalline lens and eye of the octopus. *Vision Res.* **39**, 2841–2852 (1999).

17. Fletcher, A., Murphy, T. & Young, A. Solutions of two optical problems. *Proc. R. Soc. Lond. A* **223**, 216–225 (1954).
18. Stange, G. The ocellar component of flight equilibrium control in dragonflies. *J. Comp. Physiol. A* **141**, 335–347 (1981).
19. DeAngelis, G. C., Ghose, G. M., Ohzawa, I. & Freeman, R. D. Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons. *J. Neurosci.* **19**, 4046–4064 (1999).
20. Suder, K. *et al.* Spatial dynamics of receptive fields in cat primary visual cortex related to the temporal structure of thalamocortical feedforward activity. *Exp. Brain Res.* **144**, 430–444 (2002).
21. Schweigart, G. & Eysel, U. T. Activity dependent receptive field changes in the surround of adult cat visual cortex lesions. *Eur. J. Neurosci.* **15**, 1585–1596 (2002).
22. Bartels, A. & Zeki, S. The theory of multistage integration in the visual brain. *Proc. R. Soc. Lond. B* **265**, 2327–2332 (1998).
23. Nilsson, D.-E., Andersson, M., Hallberg, E. & McIntyre, P. A micro-interferometric method for analysis of rotation-symmetric refractive-index gradients in intact objects. *J. Microsc.* **132**, 21–29 (1983).
24. Warrant, E. J. & Nilsson, D.-E. Absorption of white light in photoreceptors. *Vision Res.* **38**, 195–207 (1998).

Acknowledgements We thank E. J. Warrant and M. F. Land for comments on the manuscript, and Rita Wallén for technical assistance. This work was supported by grants from the Swedish Research Council (to D.-E.N.), the National Science Foundation USA (to M.M.C.) and the Danish Research Council (to A.G.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.-E.N. (dan-e.nilsson@cob.lu.se).

The flight paths of honeybees recruited by the waggle dance

J. R. Riley¹, U. Greggers², A. D. Smith¹, D. R. Reynolds³ & R. Menzel²

¹Plant and Invertebrate Ecology Division, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK

²Freie Universität Berlin, Fachbereich Biologie, Chemie, Pharmazie, Institut für Biologie - Neurobiologie, 28–30 Königin-Luise-Strasse, D-14195, Berlin, Germany

³Plant, Animal and Human Health Group, Natural Resources Institute, University of Greenwich, Central Avenue, Chatham, Kent ME4 4TB, UK

In the ‘dance language’ of honeybees^{1,2}, the dancer generates a specific, coded message that describes the direction and distance from the hive of a new food source, and this message is displaced in both space and time from the dancer’s discovery of that source. Karl von Frisch concluded that bees ‘recruited’ by this dance used the information encoded in it to guide them directly to the remote food source, and this Nobel Prize-winning discovery revealed the most sophisticated example of non-primate communication that we know of^{3,4}. In spite of some initial scepticism^{5–9}, almost all biologists are now convinced that von Frisch was correct^{3,4,10–14}, but what has hitherto been lacking is a quantitative description of how effectively recruits translate the code in the dance into flight to their destinations. Using harmonic radar^{15–17} to record the actual flight paths of recruited bees, we now provide that description.

There have been some important advances since von Frisch’s original experiments, and it has been believed for some time that the honeybee communication system does not instantly specify a food location to the recruits, without any hesitancy and with pinpoint accuracy¹⁸. In fact, detailed observations have shown that recruits may go through several iterations of dance session and resultant search flight before they eventually arrive at the indicated food source, and some never find the food at all^{3,19–21}. The current interpretation of the von Frisch hypothesis thus predicts that newly recruited bees should fly directly from the hive to the vicinity of a food source, and then proceed to search for its exact location using odour and other cues². This searching period would neatly account for the fact that the arrival of recruits at the source is often

very much later than would be expected for a direct flight between the hive and food source—the anomaly that caused the initial scepticism about the hypothesis.

We began our experiments by capturing recruited bees as they left our observation hive, attaching a harmonic transponder, and then releasing them. The hive was not equipped to make quantitative measurements of individual dance angles or frequencies, but we observed that, at any given time, the dances of bees that had been previously recorded at the feeder were almost all aligned about a common direction. At midday this direction was at about 90° to the vertical, and this confirmed our expectation that the bees were feeding exclusively at our artificial feeder that lay 200 m directly east of the hive, and that it was to this food source that the waggle dances referred. Neither the feed nor the feeder station itself carried any artificial scents at any time.

Most recruited bees released from the hive almost immediately undertook a straight flight of direction and length that brought them directly into the vicinity of the feeder, as shown by the 19 flight paths from the hive in the upper part of Fig. 1. Figure 2a demonstrates that the mean direction of these flights lay impressively close to the hive-to-feeder direction for the first 200 m or so but, in spite of this, only two of the recruits actually found and alighted on the feeder. This was not a surprising result, given that similar low rates of success in finding unscented feeders have been found in conventional studies of recruit flight²². At the end of their straight outward flights, some recruits promptly initiated fairly direct return flights to the hive, but more usually they engaged in what appeared to be local searching manoeuvres for several minutes. It was noticeable that searching bees occasionally passed within just a few metres of the feeder without finding it, and in a few cases, searches lasted as long as 20 min, taking recruits as far as 200 m from the feeder location. These bees often returned to the point at which they had begun their searching behaviour before flying back to the hive, but fairly direct homeward flights from other points were also seen.

In a parallel experiment, recruited bees captured leaving the hive were taken to three release points 200–250 m away, and then

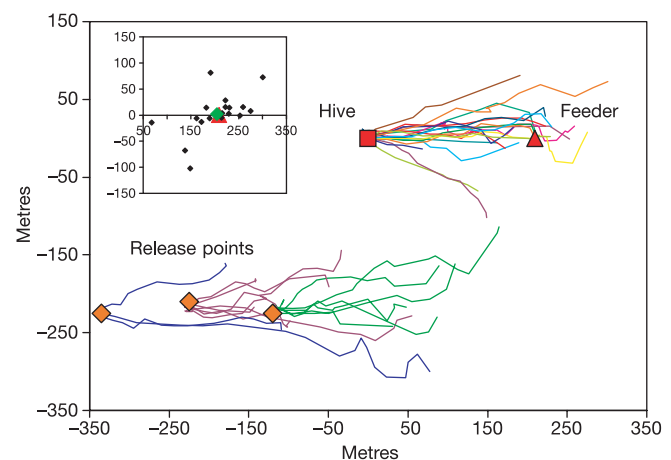


Figure 1 Flight paths of bees leaving the hive or other release points, after they had followed a waggle dance. Initial flight paths of 19 recruits leaving the hive (individual tracks in different colours) and of 17 recruits taken to locations southwest of the hive (different colour for each release point) before being released. The majority flew very close to the hive-to-feeder direction (90°), irrespective of where they were released. The tracks were truncated where they deviated by more than 90° from the hive-to-feeder direction over a track distance of more than 8 m, and this gave a mean length of 206 ± 53 m (s.d.) for hive-released bees, and 188 ± 94 m for the displaced bees. The corresponding straightness ratios (start–end point separation/track length) were 0.89 ± 0.07 and 0.82 ± 0.08 . The end-point scatter of the hive-released recruits (inset) illustrates the imprecision of the dance communication/flight navigation system, but the close proximity of the mean end point (green diamond) to the feeder (red triangle) shows that, on average, the system is remarkably accurate.

released. These bees did not fly towards the actual position of the feeder, but instead made straight flights on the vector that would have taken them to the feeder had they not been displaced (lower tracks in Fig. 1, and mean direction in Fig. 2b). Their flights were thus exactly like those of recruits released at the hive, and similarly often extended to searching behaviour at the expected feeder position. Unable to locate the feeder, the bees then returned to their release site. The processes by which they eventually returned to their hive are described elsewhere²³.

The wind fields in which the bees were flying were almost always uniform, with no evidence of large-scale mixing eddies or back flow (see the example in Fig. 3a). The flow was predominantly from the west or southwest during our experiments (Fig. 3b), typically at $2\text{--}4\text{ m s}^{-1}$, and this guaranteed that no odours from the feeder could possibly have been available to guide the recruits leaving the hive. As one would expect in these circumstances, the flight paths showed no evidence of the casting or upwind zigzagging that is known to characterize odour-following flight in other insects^{24–26}. These two facts, and the flight path data shown in Figs 1 and 2a, together carry the very convincing inference that recruits leave the hive with prior knowledge of the direction and distance of a food source that they have never previously visited. By themselves, these observations thus provide extremely strong, direct support for von Frisch's hypothesis that recruited bees 'read' the waggle dance, but the most compelling evidence comes from the flights of displaced recruits. These bees made direct flights to where the feeder should have been (Figs 1 and 2b), and did so without the aid of

any landscape or odour cues that might have existed on the true hive-to-feeder route. There was also, of course, no possibility that they were following either regular foragers directly, or ephemeral odour trails left in the hive-to-feeder flight corridor by regular forager traffic.

It is important to note that, although the feeder was in the overall downwind direction from the hive, the recruits were not simply flying downwind. For example, Fig. 3a shows how a feeder-bound recruit adopted a heading of 28° to the south of its track direction so as to compensate for the northerly drift that would otherwise have been produced by the southwesterly wind. This ability to compensate for lateral wind drift appears to be common to both bumble bees²⁷ and honeybees²⁸ and it was clearly demonstrated in most of the recruits that we observed (see examples in Fig. 3b). In spite of wind compensation, the translation by individual recruits of information encoded in the waggle dance into flight to the intended destination was rarely perfect (see scatter diagram inset, Fig. 1). While the diagram shows that the mean of the outward vector flights was impressively close to the target (within 6 m), most ended tens of metres from the feeder. Thus some form of terminal guidance would normally be essential for successful homing on the target and, in the case of natural food sources, this would usually be available in the form of visual and odour cues, and sometimes from experienced foragers manoeuvring over the source²².

In summary, we have provided the first quantitative and direct description of the degree to which recruits translate information encoded in the waggle dance into flight to the vicinity of the designated destination. Our results have shown that although this

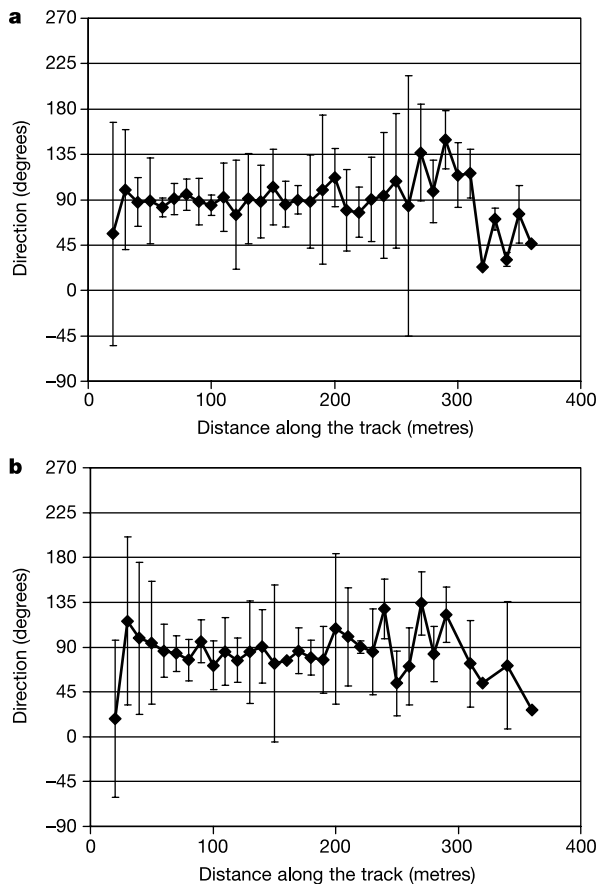


Figure 2 Circular mean of direction of travel as a function of distance along flight path. **a**, The 19 hive-departing recruits (see Fig. 1). The mean remains very close to the hive-to-feeder direction (90°) to beyond 200 m. **b**, The 17 recruits displaced to the south-west before release. The mean is again very close to 90° : the expected direction to the feeder, had the recruits read the dance. Vertical bars indicate the angular dispersion (spread of directions among individuals).

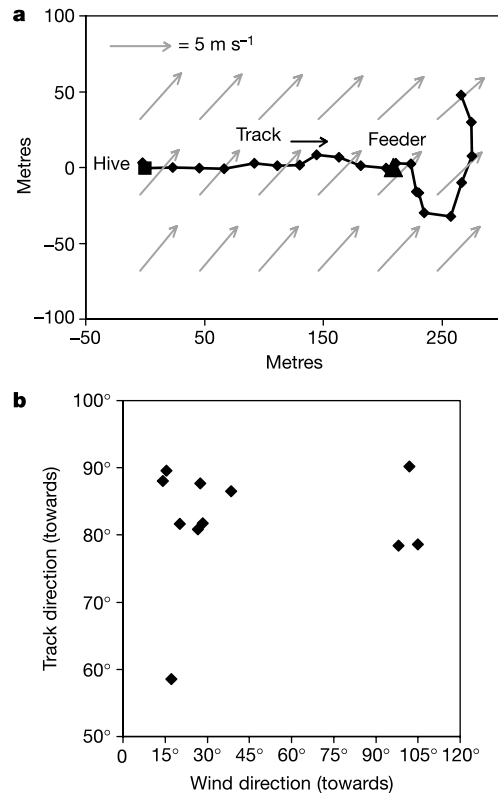


Figure 3 Wind compensation. **a**, Flight path of a recruit leaving the hive in a cross-wind (grey arrows). Over the straight part of the track (towards 87°), its ground speed was 6.8 m s^{-1} , implying a flight height of 1.9 m (refs 16, 27). The mean wind at this height was 3.3 m s^{-1} towards 38° , showing that the bee's air speed was 5.2 m s^{-1} and that it had adopted a heading of 115° to avoid drifting northwards. **b**, Comparison of wind and track directions for 11 recruits leaving the hive, when the wind field was particularly uniform. All compensated for lateral wind drift. Note also that the hive was never downwind of the feeder.

process is highly effective, most recruits would not reach the intended food sources without the use of odour and visual cues in the final stages of their flight. We hope that together with earlier studies, particularly those of Gould¹⁰, Srinivasan *et al.*¹³ and Esch *et al.*¹⁴, our results will also be accepted as a vindication of the von Frisch hypothesis. □

Methods

Harmonic radar

Most of our knowledge of insect flight behaviour at high altitudes has been derived from the use of radar modified for entomological observations²⁹. More recently, it has become possible to apply this technique to low-flying insects by tagging them with tiny harmonic transponders, weighing only a few milligrams (6–20 mg, depending on the degree of mechanical robustness required). The transponders return signals to the radar at twice the original transmitted frequency, and because these signals can be distinguished from returns from ground features and other unwanted targets, the position of tagged insects can be determined while they are in flight^{15–17}. In our experiments we caught bees as they exited their hive, attached transponders, and released them from the hive (or from remote release points), and recorded their subsequent flight trajectories. Nineteen of the 23 recruits (83%) released from the hive flew off satisfactorily, and produced good radar tracks, all to the east. Of the remaining four, three also went east, but were not detected by the radar enough times to produce a satisfactory track, probably because they were flying very low. Only one failed to leave the vicinity of the hive. Similar results were achieved for the remote releases, except during a period when the tubes in which the recruits were being transported became accidentally contaminated with sucrose solution, with the result that these bees did not fly away from the release point.

Experimental arena

The flight observations were made over a carefully selected³⁰, large area of mown pastureland, approximately 1×1.5 km, where the terrain was unusually flat and free from obstacles that would have obscured the radar's field of view²⁸. The radar was positioned on the southern edge of the arena, so that it overlooked an observation hive and a feeding station 200 m to the east of the hive. Three release points were set 200–250 m in the sector to the southwest of the hive. There were very few natural sources of pollen and nectar present during our study period (late July/early August 2000).

Description of the wind field

Wind speed and direction were recorded at 10-s intervals at a height of 2.7 m by anemometers and wind vanes placed at the corners of a 500 m $\times$ 600 m rectangle centred on the hive. We also set up a mast near to the centre of the rectangle, holding anemometers at heights of 0.65, 1.3, 2.7 and 8.2 m, and a wind vane at 2.7 m. The clocks of the recording data loggers were synchronized each morning with a master clock at the radar to ensure that wind data were recoverable for the duration of each individual recorded flight. Using interpolation methods described elsewhere^{16,27} the data collected by these instruments were then combined to describe the mean wind field within which each of the flights recorded by the radar took place.

The observation hive

We used a two-frame colony, equipped with a transparent side panel that faced directly into a small, low tent attached to the hive. From within this darkened enclosure we could observe the dance behaviour of the bees, and their entry into and exit from the hive. Most of the bees in the small colony were marked with numbered tags. The entrance to the hive was in the form of clear plastic tube, so that observers stationed outside the tent could also observe entry and exit, and capture selected bees for tagging with transponders.

The flight experiments

We began our study by establishing in our experimental bees (European species *Apis mellifera carnica*) a route memory of the position of a feeding station relative to the hive. The feeder was placed directly to the east of the hive, and supplied with 0.2–1 M sucrose solution, and training to a distance of 200 m was accomplished over two days. No artificial odour cues were used in the procedure. From the start, observers at the feeder recorded the identification number of every marked bee that arrived there, and the hive was never opened unless an observer was at the feeder. Once foraging flights between hive and feeder were well established, observers in the tent watched the waggle dances. Whenever a numbered bee was seen to follow a dance and then move directly towards the exit, an observer outside the tent was alerted to catch the bee as it attempted to leave. If its identification number indicated that it had never previously visited the feeder, the bee was confirmed as a recruit, and a transponder attached. The bee was then either released directly from the hive exit, or taken in an opaque tube to one of three release points 200–250 m from the hive, and allowed to fly from there. Bees fitted with transponders could be detected while in flight within a 190° arc of radius 900 m, centred on the radar; their positions were shown once every 3 s on the screen of a desktop personal computer, and their coordinates recorded¹⁷.

Received 28 October 2004; accepted 8 March 2005; doi:10.1038/nature03526.

1. von Frisch, K. *The Dance Language and Orientation of Bees* (Harvard Univ. Press, Cambridge, Massachusetts, 1967).
2. Dyer, F. C. The biology of the dance language. *Annu. Rev. Entomol.* **47**, 917–949 (2002).
3. Gould, J. L. Honey bee recruitment: the dance-language controversy. *Science* **189**, 685–693 (1975).
4. Sherman, G. & Visscher, P. K. Honeybee colonies achieve fitness through dancing. *Nature* **419**, 920–922 (2002).

5. Wenner, A. M. & Johnson, D. L. Honeybees: do they use direction and distance information provided by their dancers? *Science* **158**, 1076–1077 (1967).
6. Wells, P. H. & Wenner, A. M. Do honey bees have a language? *Nature* **241**, 171–175 (1973).
7. Wenner, A. M. & Wells, P. H. *Anatomy of a Controversy: the Question of a "Language" Among Bees* (Columbia Univ. Press, New York, 1990).
8. Wenner, A. M., Meade, D. E. & Friesen, L. J. Recruitment, search behavior, and flight ranges of honey bees. *Am. Zool.* **31**, 768–782 (1991).
9. Wenner, A. M. The elusive honey bee dance "language" hypothesis. *J. Insect Behav.* **15**, 859–878 (2002).
10. Gould, J. L. The dance-language controversy. *Q. Rev. Biol.* **51**, 211–244 (1976).
11. Dyer, F. C. & Smith, B. H. A review of "Anatomy of a Controversy: the Question of a 'Language' Among Bees" by A.M. Wenner & P.H. Wells. *Anim. Behav.* **47**, 1242–1244 (1994).
12. Seeley, T. D. Bee warned. *Nature* **349**, 114 (1991).
13. Srinivasan, M. V., Zhang, S. W., Altwein, M. & Tautz, J. Honeybee navigation: nature and calibration of the "odometer". *Science* **287**, 851–853 (2000).
14. Esch, H. E., Zhang, S. W., Srinivasan, M. V. & Tautz, J. Honeybee dances communicate distances measured by optic flow. *Nature* **411**, 581–583 (2001).
15. Riley, J. R. *et al.* Tracking bees with harmonic radar. *Nature* **379**, 29–30 (1996).
16. Riley, J. R. & Osborne, J. L. in *Insect Movement: Mechanisms and Consequences* (eds Woilwod, I. P., Reynolds, D. R. & Thomas, C. D.) 129–157 (CAB International, Wallingford, UK, 2001).
17. Riley, J. R. & Smith, A. D. Design considerations for an harmonic radar to investigate the flight of insects at low altitude. *Comput. Electron. Agric.* **35**, 151–169 (2002).
18. Towne, W. F. & Gould, J. L. The spatial precision of the honey bee's dance communication. *J. Insect Behav.* **1**, 129–155 (1988).
19. Esch, H. & Bastian, J. A. How do newly recruited honeybees approach a food site? *Z. Vergl. Physiol.* **68**, 175–181 (1970).
20. Mautz, D. Der Kommunikationseffekt der Schwänzeltänze bei *Apis mellifica carnica* (Pollm.). *Z. Vergl. Physiol.* **72**, 197–220 (1971).
21. Seeley, T. D. Division of labor between scouts and recruits in honeybee foraging. *Behav. Ecol. Sociobiol.* **12**, 253–259 (1983).
22. Tautz, J. & Sandeman, D. C. Recruitment of honeybees to non-scented food sources. *J. Comp. Physiol. A* **189**, 293–300 (2003).
23. Menzel, R. *et al.* Honeybees navigate according to a map-like spatial memory. *Proc. Natl Acad. Sci. USA* **102**(8), 3040–3045 (2005).
24. Baker, T. C. & Haynes, K. F. Pheromone-mediated optomotor anemotaxis and altitude control exhibited by male oriental fruit moths in the field. *Physiol. Entomol.* **21**, 20–31 (1996).
25. Cardé, R. T. & Mafra-Neto, A. in *Insect Pheromone Research: New Directions* (eds Cardé, R. T. & Minks, A. K.) 275–290 (Chapman & Hall, New York, 1997).
26. Riley, J. R. *et al.* Harmonic radar as a means of tracking the pheromone-finding and pheromone-following flight of male moths. *J. Insect Behav.* **11**, 287–296 (1998).
27. Riley, J. R. *et al.* Compensation for wind drift by bumble-bees. *Nature* **400**, 126 (1999).
28. Riley, J. R. *et al.* The automatic pilot of honeybees. *Proc. R. Soc. Lond. B* **270**, 2421–2424 (2003).
29. Reynolds, D. R. *et al.* in *Methods in Ecological and Agricultural Entomology* (eds Dent, D. R. & Walton, M. P.) 111–145 (CAB International, Wallingford, UK, 1997).
30. Chittka, L. & Geiger, K. Honeybee long-distance orientation in a controlled environment. *Ethology* **99**, 117–126 (1995).

Acknowledgements Our field studies depended on the contributions of E. Schüttler, S. Watzl, G. Bundrock, J. Stindt, S. Berger, S. Hülse, S. Brunke and T. Plümpe, and we are also grateful to A. Edwards for assistance with the reduction of the radar data and manuscript preparation. Data analysis and writing up by J.R.R. was supported by the Leverhulme Trust Emeritus Fellowship programme. The research was carried out with joint funding from the Deutsche Forschungsgemeinschaft and the Biotechnology and Agricultural Sciences Research Council of the United Kingdom (BBSRC). Rothamsted Research receives grant aided assistance from the BBSRC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.R. (joe@radarent.freeserve.co.uk).

The origin of bursts and heavy tails in human dynamics

Albert-László Barabási

Center for Complex Networks Research and Department of Physics, University of Notre Dame, Indiana 46556, USA

The dynamics of many social, technological and economic phenomena are driven by individual human actions, turning the quantitative understanding of human behaviour into a central question of modern science. Current models of human dynamics, used from risk assessment to communications, assume that human actions are randomly distributed in time and thus

well approximated by Poisson processes^{1–3}. In contrast, there is increasing evidence that the timing of many human activities, ranging from communication to entertainment and work patterns, follow non-Poisson statistics, characterized by bursts of rapidly occurring events separated by long periods of inactivity^{4–8}. Here I show that the bursty nature of human behaviour is a consequence of a decision-based queuing process^{9,10}: when individuals execute tasks based on some perceived priority, the timing of the tasks will be heavy tailed, with most tasks being rapidly executed, whereas a few experience very long waiting times. In contrast, random or priority blind execution is well approximated by uniform inter-event statistics. These findings have important implications, ranging from resource management to service allocation, in both communications and retail.

Humans participate on a daily basis in a large number of distinct activities, ranging from electronic communication (such as sending e-mails or making telephone calls) to browsing the Internet, initiating financial transactions, or engaging in entertainment and sports. Given the number of factors that determine the timing of each action, ranging from work and sleep patterns to resource availability, it seems impossible to seek regularities in human dynamics, apart from the obvious daily and seasonal periodicities. Therefore, in contrast with the accurate predictive tools common in

physical sciences, forecasting human and social patterns remains a difficult and often elusive goal.

Current models of human activity are based on Poisson processes, and assume that in a dt time interval an individual (agent) engages in a specific action with probability qdt , where q is the overall frequency of the monitored activity. This model predicts that the time interval between two consecutive actions by the same individual, called the waiting or inter-event time, follows an exponential distribution (Fig. 1a–c)¹. Poisson processes are widely used to quantify the consequences of human actions, such as modelling traffic flow patterns or accident frequencies¹, and are commercially used in call centre staffing², inventory control³, or to estimate the number of congestion-caused blocked calls in calls in mobile communication⁴. Yet, an increasing number of recent measurements indicate that the timing of many human actions systematically deviates from the Poisson prediction, the waiting or inter-event times being better approximated by a heavy tailed or Pareto distribution (Fig. 1d–f). The differences between Poisson and heavy-tailed behaviour are striking: a Poisson distribution decreases exponentially, forcing the consecutive events to follow each other at relatively regular time intervals and forbidding very long waiting times. In contrast, the slowly decaying, heavy-tailed processes allow for very long periods of inactivity that separate bursts of intensive activity (Fig. 1).

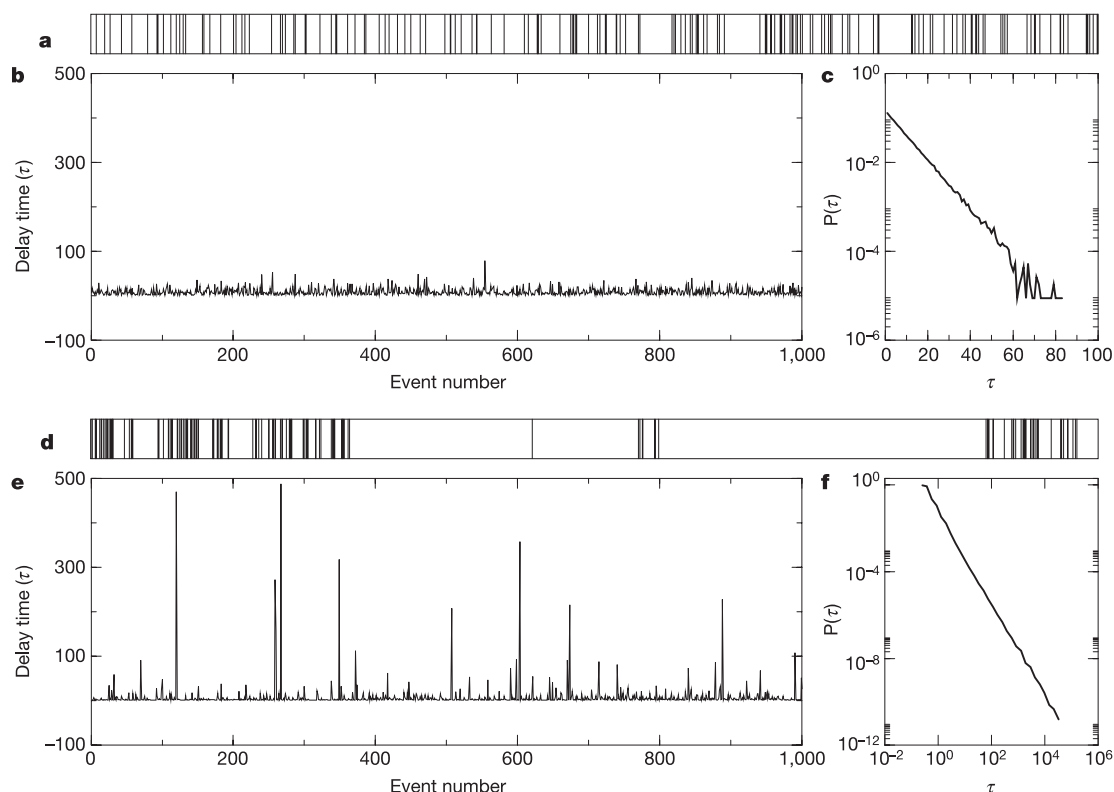


Figure 1 The difference between the activity patterns predicted by a Poisson process and the heavy-tailed distributions observed in human dynamics. **a**, Succession of events predicted by a Poisson process, which assumes that in any moment an event takes place with probability q . The horizontal axis denotes time, each vertical line corresponding to an individual event. Note that the inter-event times are comparable to each other, long delays being virtually absent. **b**, The absence of long delays is visible on the plot showing the delay times τ for 1,000 consecutive events, the size of each vertical line corresponding to the gaps seen in **a**. **c**, The probability of finding exactly n events within a fixed time interval is $P(n, q) = e^{-qt}(qt)^n/n!$, which predicts that for a Poisson process the inter-event time distribution follows $P(\tau) = qe^{-q\tau}$, shown on a log-linear plot in **c** for the

events displayed in **a**, **b**, **d**. The succession of events for a heavy-tailed distribution.

e, The waiting time τ of 1,000 consecutive events, where the mean event time was chosen to coincide with the mean event time of the Poisson process shown in **a–c**. Note the large spikes in the plot, corresponding to very long delay times. **b** and **e** have the same vertical scale, allowing the comparison of the regularity of a Poisson process with the intermittent nature of the heavy-tailed process. **f**, Delay time distribution $P(\tau) \approx \tau^{-2}$ for the heavy-tailed process shown in **d**, **e**, appearing as a straight line with slope -2 on a log-log plot. The signal shown in **d–f** was generated using $\gamma = 1$ in the stochastic priority list model discussed in the Supplementary Information.

To provide direct evidence for non-Poisson activity patterns in individual human behaviour, I study the communication between several thousand e-mail users based on a data set capturing the sender, recipient, time and size of each e-mail^{11–12}. As Fig. 2a shows, the distribution of the time differences between consecutive e-mails sent by a selected user is best approximated with $P(\tau) \approx \tau^{-\alpha}$, where $\alpha \approx 1$, indicating that an individual's e-mail pattern has a bursty non-Poisson character: during a single session a user sends several e-mails in quick succession, followed by long periods of no e-mail activity. This behaviour is not limited to e-mail communications. Measurements capturing the distribution of the time differences between consecutive instant messages sent by individuals during online discussions⁵ show a similar pattern. Professional tasks, such as the timing of job submissions on a supercomputer⁶, directory listing and file transfers (FTP request) initiated by individual users⁷, or the timing of printing jobs submitted by users¹³ were also reported to display non-Poisson features. Similar patterns emerge in economic transactions, describing the time interval distributions between individual trades in currency futures⁸. Finally, heavy-tailed distributions characterize entertainment-related events, such as the time intervals between consecu-

tive online games played by the same user¹⁴.

The fact that a wide range of human activity patterns follow non-Poisson statistics suggests that the observed bursty character reflects some fundamental and potentially generic feature of human dynamics. Yet, the mechanism responsible for these marked non-random features remains unknown. Here, I show that the bursty nature of human dynamics is a consequence of a queuing process driven by human decision making: whenever an individual is presented with multiple tasks and chooses among them based on some perceived priority parameter, the waiting time of the various tasks will follow a Pareto distribution. In contrast, first-come-first-serve and random task execution, common in most service-oriented or computer-driven environments, lead to uniform Poisson-like dynamics.

Most human-initiated events require an individual to assess and prioritize different activities. Indeed, at the end of each activity an individual needs to decide what to do next—for example send an e-mail, do some shopping, or make a telephone call—allocating time and resources for the chosen activity. Consider an agent operating with a priority list of L tasks. After a task is executed, it is removed from the list, offering the opportunity to add another task. The agent assigns to each task a priority parameter x , which allows it to compare the urgency of the different tasks on the list. The question is, how long will a given task have to wait before it is executed. The answer depends on the method the agent uses to choose the task to be executed next. In this respect three selection protocols¹⁰ are particularly relevant for human dynamics.

(i) The simplest selection rule is the first-in-first-out protocol, executing the tasks in the order that they were added to the list. This is common in service-oriented process, such as the first-come-first-serve execution of orders in a restaurant or getting help from directory assistance and consumer support. The time period an item stays on the list before execution is determined by the cumulative time required to perform all tasks added to the list before it. If the time necessary to perform the individual tasks are chosen from a bounded distribution (that is, the second moment of the distribution is finite), then the waiting time distribution will develop an exponential tail, indicating that most tasks experience approximately the same waiting time.

(ii) The second possibility is to execute the tasks in a random order, irrespective of their priority or time spent on the list. This is common, for example, in educational settings, when students are called on randomly, and in some packet routing protocols in Internet communications. The waiting time distribution of the individual tasks (that is, the time between two calls on the same student) in this case is also exponential.

(iii) In most human-initiated activities task selection is not random, but the individual executes the highest-priority item on its list. The resulting execution dynamics is quite different from the first (i) and second (ii) selection protocols: high-priority tasks will be executed soon after their addition to the list, whereas low-priority items will have to wait until all higher-priority tasks are cleared, forcing them to stay on the list for considerable time intervals. Below, I show that this selection mechanism, practiced by humans on a daily basis, is the probable source of the fat tails observed in human-initiated processes.

I assume that an individual has a priority list with L tasks, each task being assigned a priority parameter x_i , where $i = 1, \dots, L$, chosen from a $\rho(x)$ distribution. At each time step the agent selects the highest-priority task from the list and executes it, removing it from the list. At that moment a new task is added to the list, its priority x_i being again chosen from $\rho(x)$. This simple model ignores the possibility that the agent occasionally selects a low-priority item for execution before all higher-priority items are done—common, for example, for tasks with deadlines. This can be incorporated by assuming that the agent executes the highest-priority item with probability p , and with probability $1 - p$ executes a randomly

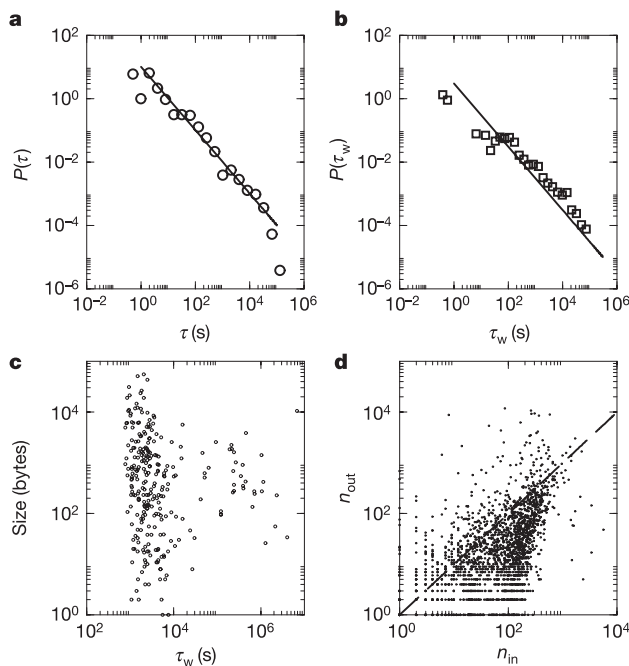


Figure 2 Heavy-tailed activity patterns in e-mail communications. **a**, The distribution of the time intervals between consecutive e-mails sent by a single user over a three-month time interval, indicating that $P(\tau) \approx \tau^{-1}$ (the solid line in the log-log plot has slope -1). Although the exponent differs slightly from user to user, it is typically centred around $\alpha = 1$. **b**, The distribution of the time taken by the user to reply to a received message. To determine τ_w we recorded the time the user received an e-mail from a specific user, and the time it sent a response to that user, the difference between the two providing τ_w . For consistency the figure shows the data for the user whose inter-event time distribution is shown in **a**. The solid line in the log-log plot has slope -1 . **c**, A scatter plot showing the waiting time τ_w and the size for each e-mail responded to by the user discussed in **a**, **b**, indicating that the file size and response time do not correlate. **d**, Scatter plot showing the number of e-mails received and sent by 3,188 users during a three-month interval. Each point corresponds to a different user, indicating that there are significant differences between the number of received and responded e-mails. The dashed line corresponds to $n_{in} = n_{out}$, capturing the case when the classical queuing models also predict a power law waiting time distribution (see Supplementary Information), albeit with exponent $\alpha = 3/2$.

selected task, independent of its priority. Thus the $p \rightarrow 1$ limit of the model describes the deterministic protocol (iii), when always the highest-priority task is chosen for execution whereas $p \rightarrow 0$ corresponds to the random choice protocol (ii) discussed above.

To establish that this priority list model can account for the observed fat-tailed inter-event time distribution, we first studied its dynamics numerically with priorities chosen from a uniform distribution $x_i \in [0, 1]$. Computer simulations show that in the $p \rightarrow 1$ limit the probability that a task spends τ time on the list has a power law tail with exponent $\alpha = 1$ (Fig. 3a) in agreement with the exponent obtained for e-mail communications (Fig. 2a). In the $p \rightarrow 0$ limit $P(\tau)$ follows an exponential distribution (Fig. 3b), as expected for the case (ii). As the typical length of the priority list differs from individual to individual, it is particularly important for the tail of $P(\tau)$ to be independent of L . Numerical simulations indicate that this is indeed the case: changes in L do not affect the scaling of $P(\tau)$. The fact that the scaling holds for $L = 2$ indicates that it is not necessary to have a long priority list: as long as individuals balance at least two tasks, a bursty, heavy-tailed inter-event dynamics will emerge.

To determine the tail of $P(\tau)$ analytically I consider a stochastic version of the model in which the probability to choose a task with priority x for execution in a unit of time is $\Pi(x) \approx x^\gamma$, where γ is a parameter that allows us to interpolate between the random choice limit (ii, $\gamma = 0$, $p = 0$) and the deterministic case, when always the highest-priority item is chosen for execution (iii, $\gamma = \infty$, $p = 1$). Note that this parameterization captures the scaling of the model only in the $p \rightarrow 0$ and $p \rightarrow 1$ limits, but not for intermediate p values, thus it is chosen only for mathematical convenience. The probability that a task with priority x is executed at time t is $f(x, t) = (1 - \Pi(x))^{t-1} \Pi(x)$. The average waiting time of a task with priority x is obtained by averaging over t weighted with $f(x, t)$ providing

$$\tau(x) = \sum_{t=1}^{\infty} t f(x, t) = \frac{1}{\Pi(x)} \approx \frac{1}{x^\gamma} \quad (1)$$

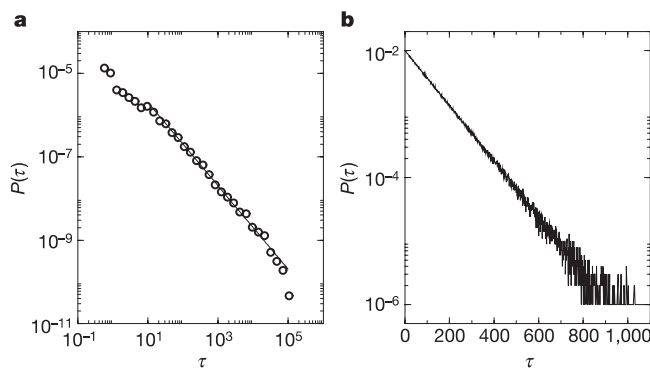


Figure 3 The waiting time distribution predicted by the investigated queuing model. The priorities were chosen from a uniform distribution $x_i \in [0, 1]$, and I monitored a priority list of length $L = 100$ over $T = 10^6$ time steps. **a**, Log-log plot of the tail of probability $P(\tau)$ that a task spends τ time on the list obtained for $p = 0.99999$, corresponding to the deterministic limit of the model. The continuous line of the log-log plot corresponds to the scaling predicted by equation (2), having slope -1 , in agreement with the numerical results and the analytical predictions. The data were log-binned, to reduce the uneven statistical fluctuations common in heavy-tailed distributions, a procedure that does not alter the slope of the tail. For the full curve, including the $\tau = 1$ peak, see Supplementary Fig. 3. **b**, Linear-log plot of the $P(\tau)$ distribution for $p = 0.00001$, corresponding to the random choice limit of the model. The fact that the curve follows a straight line on a linear-log plot indicates that $P(\tau)$ decays exponentially.

that is, the higher an item's priority, the shorter the average time it waits before execution. To calculate $P(\tau)$ I use the fact that the priorities are chosen from the $\rho(x)$ distribution; that is, $\rho(x)dx = P(\tau)d\tau$, which gives

$$P(\tau) \approx \frac{\rho(\tau^{-1/\gamma})}{\tau^{1+1/\gamma}} \quad (2)$$

In the $\gamma \rightarrow \infty$ limit, which converges to the strictly priority-based deterministic choice ($p = 1$) in the model, equation (2) predicts $P(\tau) \approx \tau^{-1}$, in agreement with the numerical results (Fig. 3a), as well as the empirical data on the e-mail inter-arrival times (Fig. 2a). In the $\gamma = 0$ ($p = 0$) limit, $\tau(x)$ is independent of x thus $P(\tau)$ converges to an exponential distribution, as shown in Fig. 3b (see Supplementary Information).

The apparent dependence of $P(\tau)$ on the $\rho(x)$ distribution from which the agent chooses the priorities may appear to represent a potential problem, as assigning priorities is a subjective process, each individual being characterized by its own $\rho(x)$ distribution. According to equation (2), however, in the $\gamma \rightarrow \infty$ limit, $P(\tau)$ is independent of $\rho(x)$. Indeed, in the deterministic limit the uniform $\rho(x)$ can be transformed into an arbitrary $\rho'(x)$ with a parameter change, without altering the order in which the tasks are executed¹¹. This insensitivity of the tail to $\rho(x)$ explains why, despite the diversity of human actions encompassing both professional and personal priorities, most decision-driven processes develop a heavy tail.

To obtain empirical evidence for the validity of the proposed queuing mechanism I consider the e-mail activity pattern of an individual^{11,12}. Once in front of a computer, an individual will reply immediately to a high-priority message, while placing the less urgent or the more difficult ones on its priority list to compete with other non-e-mail activities. I propose, therefore, that the observed inter-event time distribution is in fact rooted in the uneven waiting times experienced by different tasks. To test this hypothesis the waiting time for each task needs to be determined directly. In the e-mail data set we have the time, sender and recipient of each e-mail transmitted over several months by each user, thus we can determine the time it takes for a user to reply to a received message¹¹. As Fig. 2b shows, the waiting time distribution $P(\tau_w)$ for the user whose $P(\tau)$ is shown in Fig. 2a is best approximated by $P(\tau_w) \approx \tau_w^{-\alpha_w}$ with exponent $\alpha_w = 1$, supporting the hypothesis that the heavy-tailed waiting time distribution drives the observed bursty e-mail activity patterns.

As in the $p \rightarrow 1$ limit of the model the priority list is dominated by low-priority tasks, new tasks will often be executed immediately. This results in a peak at $P(\tau = 1)$ (see Supplementary Fig. 3), which, although in some cases may represent a model artefact, in the e-mail context is not unrealistic: most e-mails are either deleted right away (which is one kind of task execution) or immediately replied to. Only the more difficult or time consuming tasks will queue on the priority list. The e-mail data set does not allow us to resolve this peak, however, because a message which the user deletes or replies to right away will appear to have some waiting time, given the delay between the arrival of the message and the time the user has a chance to check her e-mail.

Although I have illustrated the queuing process for e-mails, in general the model is better suited to capture the competition between different kinds of activities an individual is engaged in; that is, the switching between various work, entertainment and communication events. Indeed, most data sets displaying heavy-tailed inter-event times in a specific activity reflect the outcome of the competition between tasks of different nature. For example, the starting of an online gaming session often implies that all higher-priority work-, family-, and entertainment-related activities have been already executed.

Detailed models of human activity require us to consider the impact of a number of additional mechanisms on the queuing

process. First, in the priority list model I have assumed that the time necessary to execute a task (service time) is the same for all tasks. The size distribution of e-mails is heavy tailed^{15,16}, however, thus the waiting time distribution could be driven entirely by the time it takes to read an e-mail (that is, the message size). Yet, as Fig. 2c shows, there is no correlation between the size of the e-mail received by a user and the time the user takes to reply to it. Although detailed analysis should also consider the role of attachments, Fig. 2c suggests that the priority of a response is more important than the message size. Furthermore, the priorities assigned to tasks are often driven by optimization processes, as agents aim to maximize profits or minimize the overall time spent on some activity.

A natural extension of the model is to assume that tasks arrive at a rate λ and are executed at a rate μ , allowing the length of the priority list L to change in time. In this case the model maps into Cobham's priority queue model⁹, which has a power-law-distributed waiting time with $\alpha = 3/2$ when $\lambda = \mu$ (see Supplementary Information). Thus, to account for the power law waiting times the model requires an additional mechanism that guarantees $\lambda = \mu$ (which, as Fig. 3d indicates, is not satisfied for most e-mail users). In contrast, in the proposed priority list model I have assumed that for humans the length of the priority list remains relatively unchanged (L is constant). To understand the origin of this assumption we must realize that for $\lambda = \mu$ the length of the priority list fluctuates widely and can occasionally grow very long. Although keeping track of a long priority list is not a problem for a computer, it is well established that the immediate memory of humans has finite capacity¹⁷. In other words, the number of priorities that we can easily remember, and therefore the length of the priority list, is bounded, motivating the choice of a finite L .

Although other generalizations are possible and often required, the main finding is that the observed fat-tailed activity distributions can be explained by a simple hypothesis: humans execute their tasks based on some perceived priority, setting up queues that generate very uneven waiting time distributions for different tasks. In this respect the proposed priority list model represents only a minimal framework that allows us to demonstrate the potential origin of the heavy-tailed activity patterns, and offers room for further extensions to capture more complex human behaviour. As the exponent of the tail could depend on the details of the prioritizing process, future work may allow the empirical data to discriminate between different queuing hypotheses. A mapping into punctuated equilibrium models (see Supplementary Information and refs 18, 19) with the mathematical framework of queuing theory could help the systematic classification of the various temporal patterns generated by human behaviour.

There is overwhelming evidence that Internet traffic is characterized by heavy-tailed statistics²⁰, rooted in the Pareto size distribution of the transmitted files^{15,16}. As I have shown (Fig. 2c), a user's e-mail activity does not correlate with e-mail size. Similarly, the timing of online games¹⁴ or sending an instant message⁵ cannot be driven by file size either. This suggests that Internet traffic is in fact driven by two separate processes: the heavy-tailed size distribution of the files sent by the users, and the human decision-driven timing of various Internet-mediated activities individuals engage in. In some environments this second mechanism, the origin of which is addressed in this paper, can be just as important as the much investigated first one. Given the differences in routing performance under Poisson and Pareto arrival time distributions^{20–22}, a queuing-based model of human-driven arrival times could also contribute to a better understanding of Internet traffic.

Uncovering the mechanisms governing the timing of various human activities has significant scientific and commercial potential. First, models of human behaviour are indispensable for large-scale models of social organization, ranging from detailed urban models^{23,24} to modelling the spread of epidemics and viruses, the

development of panic²⁵ or capturing financial market behaviour²⁶. Understanding the origin of the non-Poisson nature of human dynamics could fundamentally alter the dynamical conclusions these models offer. Second, models of human behaviour are crucial for better resource allocation and pricing plans for telephone companies, to improve inventory and service allocation in both online and "high street" retail, and potentially to understand the bursts of ideas and memes emerging in communication and publication patterns²⁷. Finally, heavy tails have been observed in the foraging patterns of birds as well²⁸, raising the intriguing possibility that animals also use some evolutionarily encoded priority-based queuing mechanisms to decide between competing tasks, such as caring for offspring, gathering food, or fighting off predators. □

Received 7 November 2004; accepted 15 February 2005; doi:10.1038/nature03459.

1. Haight, F. A. *Handbook of the Poisson Distribution* (Wiley, New York, 1967).
2. Reynolds, P. *Call Center Staffing* (The Call Center School Press, Lebanon, Tennessee, 2003).
3. Greene, J. H. *Production and Inventory Control Handbook* 3rd edn (McGraw-Hill, New York, 1997).
4. Anderson, H. R. *Fixed Broadband Wireless System Design* (Wiley, New York, 2003).
5. Dewes, C., Wichmann, A. & Feldman, A. in *Proc. 2003 ACM SIGCOMM Conf. Internet Measurement (IMC-03)* (ACM Press, New York, 2003).
6. Kleban, S. D. & Clearwater, S. H. Hierarchical Dynamics, Interarrival Times and Performance. *Proc. SC2003* (<http://www.sc-conference.org/sc2003/paperpdfs/pap222.pdf>) (2003).
7. Paxson, V. & Floyd, S. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Trans. Netw.* **3**, 226 (1996).
8. Masoliver, J., Montero, M. & Weiss, G. H. Continuous-time random-walk model for financial distributions. *Phys. Rev. E* **67**, 021112 (2003).
9. Cobham, A. Priority assignment in waiting line problems. *J. Oper. Res. Soc. Am.* **2**, 70–76 (1954).
10. Cohen, J. W. *The Single Server Queue* (North Holland, Amsterdam, 1969).
11. Eckmann, J.-P., Moses, E. & Sergi, D. Entropy of dialogues creates coherent structure in e-mail traffic. *Proc. Natl Acad. Sci. USA* **101**, 14333–14337 (2004).
12. Ebel, H., Mielsch, L. I. & Bornholdt, S. Scale-free topology of e-mail network. *Phys. Rev. E* **66**, R35103 (2002).
13. Harder, U. & Paczuski, M. Correlated dynamics in human printing behavior. Preprint at (<http://xxx.lanl.gov/abs/cs.PF/0412027>) (2004).
14. Henderson, T. & Nhatti, S. Modelling user behavior in networked games. *Proc. 9th ACM Int. Conf. on Multimedia* 212–220 (ACM Press, New York, 2001).
15. Crovella, M. & Bestavros, A. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Trans. Netw.* **5**, 835–846 (1997).
16. Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**, 226–251 (2004).
17. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **63**, 8197 (1956).
18. Bak, P. & Sneppen, K. Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.* **71**, 4083–4086 (1993).
19. Jensen, H. J. *Self Organised Criticality* (Cambridge Univ. Press, Cambridge, 1998).
20. Park, K. & Willinger, W. *Self-Similar Networks Traffic and Performance Evaluation* (Wiley, New York, 2000).
21. Leighton, F. T., Maggs, B. M. & Rao, S. B. Packet routing and job-shop scheduling in $O(\text{congestion} + \text{dilation})$ steps. *Combinatorica* **14**, 167–186 (1994).
22. Harris, C. M., Brill, P. H. & Fischer, M. J. Internet-type queues with power-tailed interarrival times and computational methods for their analysis. *INFORMS J. Comp.* **12**, 261–271 (2000).
23. Eubank, H. et al. Controlling epidemics in realistic urban social networks. *Nature* **429**, 180–184 (2004).
24. Manrubia, S. C., Zanette, D. & Sole, R. V. Transient dynamics and scaling phenomena in urban growth. *Fractals* **7**, 1–8 (1999).
25. Helbing, D., Farkas, I. & Vicsek, T. Simulating dynamic features of escape panic. *Nature* **407**, 487–490 (2000).
26. Caldarelli, G., Marsili, M. & Zhang, Y.-C. A prototype model of stock exchange. *Europhys. Lett.* **40**, 479–484 (1997).
27. Kleinberg, J. in *Proc. 8th ACM SIGKDD Intl Conf. Knowledge Discov. Data Mining*, 91–101 (2002).
28. Viswanathan, G. M. et al. Optimizing the success of random searches. *Nature* **401**, 911–914 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements I have benefited from discussions with A. Vazquez on the mathematical aspects of the model. I also thank L. A. N. Amaral, Z. Dezső, P. Ivanov, J. Kelley, J. Kertész, A. Motter, M. Paczuski, K. Sneppen, T. Vicsek, W. Whitt and E. Zambrano for useful discussions and comments on the manuscript; J.-P. Eckmann for providing the e-mail database; and S. Aleva for assisting me with manuscript preparation. This research was supported by NSF grants.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to A.-L. B. (alb@nd.edu).

Fast vesicle replenishment allows indefatigable signalling at the first auditory synapse

Claudius B. Griesinger^{1,3}, Christopher D. Richards¹
& Jonathan F. Ashmore^{1,2}

¹Department of Physiology and ²UCL Ear Institute, University College London, Gower Street, London WC1E 6BT, UK

³Physiologisches Institut II, Universität Freiburg, Hermann Herder Str. 7, 79104 Freiburg, Germany

Ribbon-type synapses in inner hair cells of the mammalian cochlea encode the complexity of auditory signals by fast and tonic release through fusion of neurotransmitter-containing vesicles. At any instant, only about 100 vesicles are tethered to the synaptic ribbon, and about 14 of these are docked to the plasma membrane^{1,2}, constituting the readily releasable pool³. Although this pool contains about the same number of vesicles as that of conventional synapses^{4,5}, ribbon release sites operate at rates of about two orders of magnitude higher^{3,6,7} and with submillisecond precision^{8–11}. How these sites replenish their vesicles so efficiently remains unclear^{3,12,13}. We show here, using two-photon imaging of single release sites in the intact cochlea, that preformed vesicles derived from cytoplasmic vesicle-generating compartments¹⁴ participate in fast release and replenishment. Vesicles were released at a maximal initial rate of 3 per millisecond during a depolarizing pulse, and were replenished at a rate of 1.9 per millisecond. We propose that such rapid resupply of vesicles enables temporally precise and sustained release rates. This may explain how the first auditory synapse can encode with indefatigable precision without having to rely on the slow, local endocytic vesicle cycle⁷.

Here, we have studied vesicle dynamics during release and replenishment using high-resolution two-photon imaging of individual release sites in inner hair cells (IHCs). Release was triggered by transepithelial current stimulation¹⁵ of the intact organ of Corti (Fig. 1a)¹⁴. As with other cells containing ribbon synapses¹⁶, IHCs contain an extensive pool of vesicles in their synaptic cytoplasm¹⁷. A substantial proportion of these are release-competent¹⁸ and are derived from vesicle-generating compartments in the apex, before being transported to basal clusters and release sites¹⁴. Taking advantage of this feature of IHCs, we labelled the cytoplasmic vesicle population by apical endocytosis of the fluorescent membrane marker FM1-43 (Fig. 1b). This procedure led to aggregates of fluorescent vesicles at release sites in the subnuclear zone (Fig. 1c). Because the epithelium was intact and dye could not reach basolateral membranes¹⁴, fluorescence was not derived from local basolateral endocytosis¹⁹. From cell reconstructions, the average number of release sites per cell was 22 ± 4 (17 cells in 5 cochleae) in agreement with earlier estimates of 18–25 sites per IHC^{1,20}. These sites had an average diameter of $0.79 \pm 0.30 \mu\text{m}$ ($n = 186$, 43 cells, 21 cochleae). By converting fluorescence intensity into vesicle number, we estimate¹⁴ that each site held 400 vesicles (range 45–1,600). This number probably includes vesicles in the nearby cytoplasm, because electron microscopy suggests that only 100 vesicles are directly tethered to a 200-nm diameter ribbon^{1,2}.

The IHC calcium currents had the properties of a $\text{Ca}_v1.3$ current with the α_{1D} subunit; that is, non-inactivating L-type calcium current (Supplementary Fig. 1). We therefore investigated whether L-type calcium channel blockers (either 10 μM nimodipine or 10 μM nifedipine) could affect release site de-staining. In control conditions, stimulation of IHCs by a 40-Hz train of 20-ms pulses (200- μA amplitude, 50-s duration) decreased release

site fluorescence by $31 \pm 2\%$ ($n = 7$ cells) (Fig. 1d). The fluorescence de-staining was reversibly blocked by 10 μM nimodipine (Fig. 1d) and reduced by 10 μM nifedipine in the bath (see also Supplementary Fig. 1). To exclude movement artefacts, release sites were imaged using z-stack time series with 5–7 imaging planes.

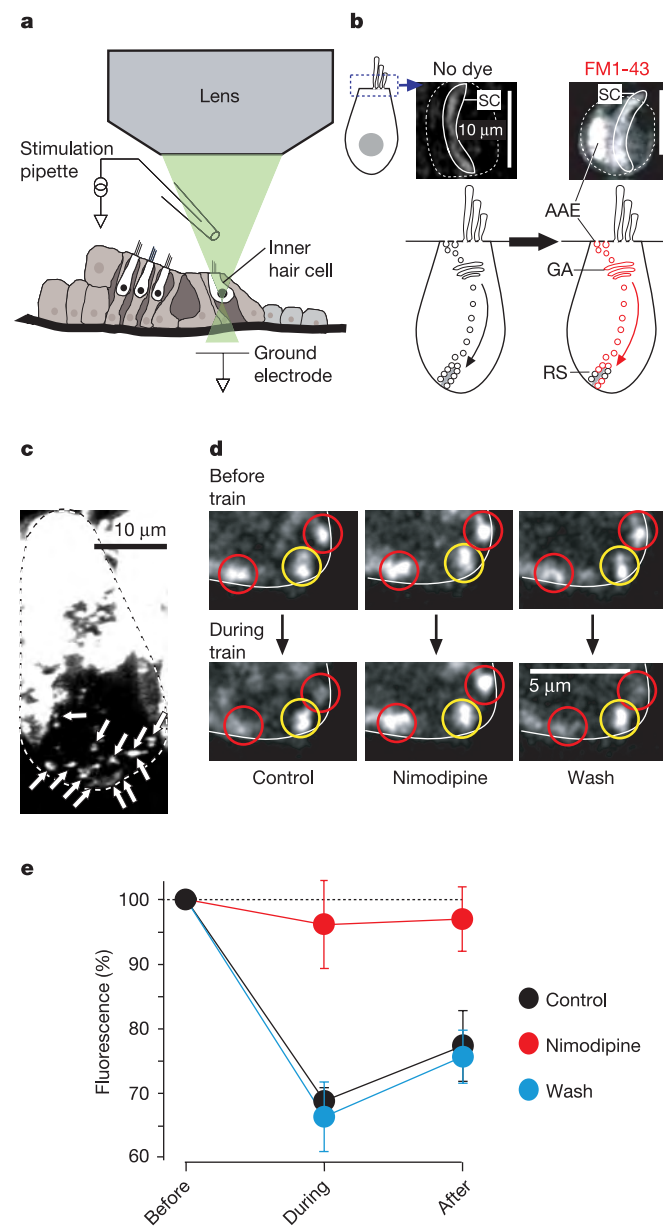


Figure 1 Experimental design. **a**, Schematic view of the experimental arrangement. **b**, FM1-43 labelling of preformed vesicle pools in the basal cytoplasm at release sites by continuous membrane transport (arrows). FM1-43 was endocytosed from the apical membrane labelling apical early endosomes (AAE) and the Golgi apparatus (GA). Labelled vesicles were transported (arrows) to basal release sites (RS). Upper images are reconstructed from z-stacks of the apex of a cell before and after FM1-43 uptake; cartoon indicates position of stack. Stereocilia (SC) were identified by autofluorescence. **c**, Release sites (arrows) in a cell labelled by apical FM1-43 uptake. **d**, Nimodipine blocks de-staining. Top panels: release sites before stimulation in control, 10 μM nimodipine and during wash. Bottom panels: the first frame after stimulus onset in these conditions. Red, responding release sites; yellow, non-responding release site. **e**, Fluorescence decrease with and without nimodipine in the bath. For each release site, fluorescence values for the baseline (before), during the stimulus train (during) and for the recovery period (after) were averaged and normalized to baseline values. These values were then averaged over ten release sites from three preparations. Error bars indicate s.d.

De-staining was therefore a measure of release of FM1-43-labelled synaptic vesicles.

At the highest temporal and spatial resolutions, de-staining of the vesicle cluster was extensive (Fig. 2a). Fluorescence, although decreasing at the ribbon, increased in the plasma membrane and nearby cytosol (Fig. 2b; see also Supplementary Fig. 2). With long stimulation trains (>10 s), fluorescence began to recover while stimulation continued (Fig. 2c, black arrow). After cessation of stimulation, fluorescence returned towards baseline with a time constant of 1.9 s (Fig. 2c, red arrow). Because endocytosis in IHCs has a time constant of 7.5 s under physiological conditions³, fluorescence recovery represents translocation of preformed and fluorescently labelled vesicles from cytoplasmic pools to release sites.

The kinetics and amplitude of de-staining were dependent on both the amplitude and frequency of the stimulus train (data not shown), consistent with de-staining representing release. By varying current amplitudes from 50 to 300 μ A, we found two populations of release sites that differ in their response threshold (data not shown). One population responded at 100 μ A and showed maximum de-staining at 200 μ A. The second had a higher de-staining threshold (>200 μ A). Here we concentrate on the first population because its full dynamic range could be studied without evoking movement artefacts.

Short pulses, comparable to those used in capacitance measurements³, lead to optically detectable release. Figure 2d shows that single pulses of 75- and 150-ms duration produced de-staining that increased exponentially with pulse length ($\tau = 30$ ms) (Fig. 2e).

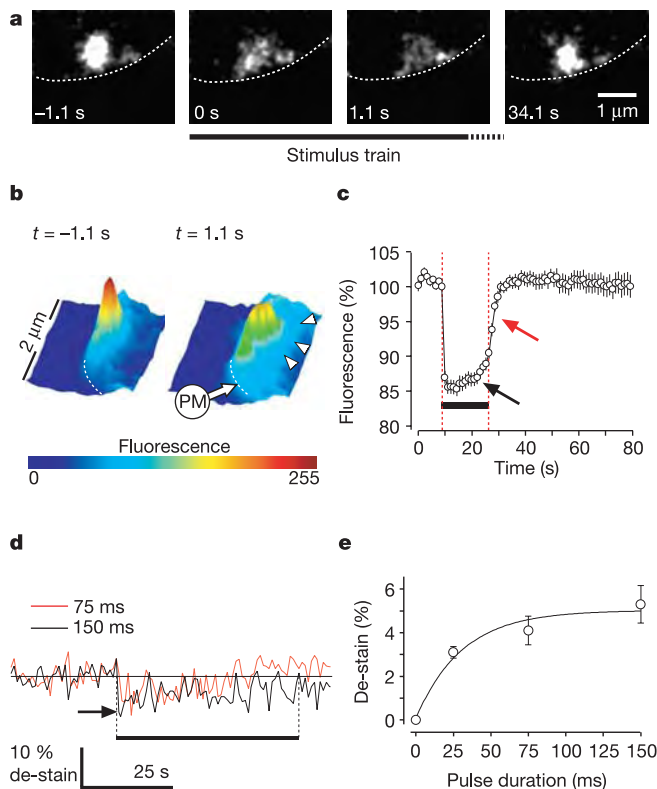


Figure 2 Imaging individual ribbon release sites. **a**, Release site before, during and after a stimulus. The stimulus consisted of a train of 750 pulses (20 ms, 200 μ A, at 40 Hz). Stimulus onset at 0 s. **b**, Intensity-contour plots of a site before and after onset of stimulation. Stimulation as in **a**. Fluorescence decreased at the release sites, but increased in the plasma membrane (PM) and in the cytosol (arrowed). **c**, Average of 54 release sites stimulated as in **a**. Error bars indicate standard error of the mean (s.e.m.). **d**, Fluorescence from one release site stimulated with 75- and 150-ms pulses at 1 Hz for 50 s (bar). One pulse elicited measurable de-staining (arrow). **e**, Relationship between (single) pulse duration and de-staining (s.e.m., $n = 9$ release sites).

For better temporal resolution, we performed line-scan imaging of individual sites (Figs 3 and 4). A total of 15 individual release sites were analysed in this manner. Figure 3a, b shows an example with four release sites in two neighbouring cells. Upon stimulation, three sites responded by de-staining 20–60%. Fluorescence loss continued throughout stimulation (sites 2 and 4). It recovered to near baseline values at the end of the pulse. Fluorescence recovery rates were at least one order of magnitude faster than reported endocytosis rates in IHCs, where time constants of 7.5 s have been determined using physiological stimuli³, whereas faster rates ($\tau = 0.3$ s) occur at cellular calcium levels around 30 μ M, which might arise locally at release sites¹⁸. In most line-scan experiments, fluorescence recovered to around 90% of the pre-stimulus value (see Figs 3b inset, e and 4c), indicating that either there is incomplete repopulation of the ribbon or a small fraction of replenishing vesicles originates from non-labelled sources. In line-scan experiments, the time course of de-staining depended on the stimulation amplitude, whereas that of recovery did not. Figure 3e shows that at five release sites, de-staining time constants were different with

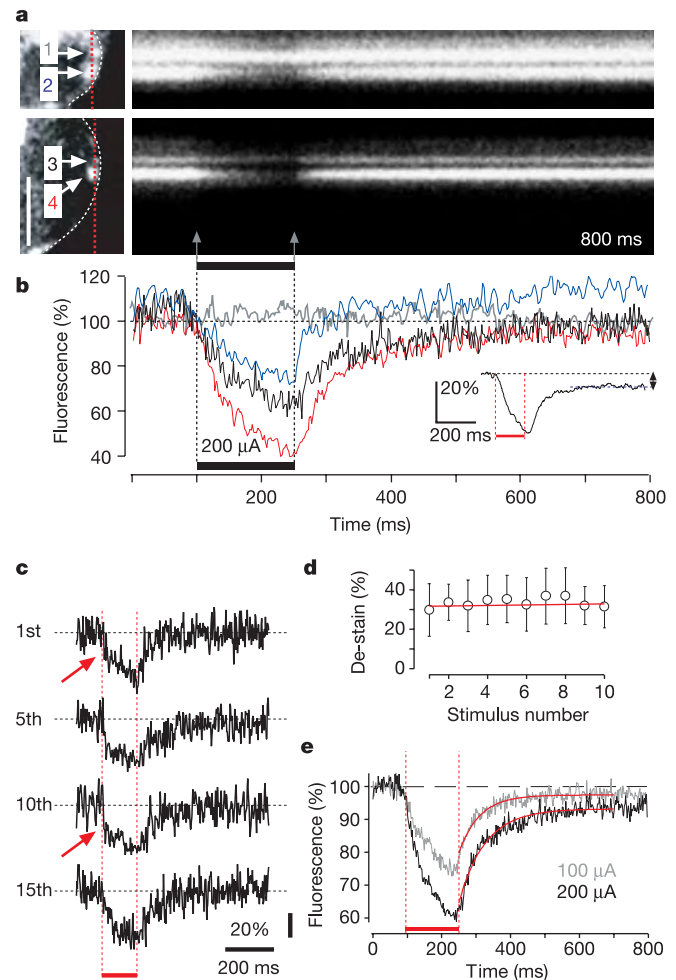


Figure 3 Kinetics of release and recovery at individual release sites. **a**, Line scan through four release sites in two neighbouring cells. Left: images indicate scan line position. Scale bar, 5 μ m. Right: average of 12 line-scan images (800 ms) of release triggered by a 150-ms, 200- μ A pulse. **b**, Average of 12 consecutive release events of the four sites in **a**. Inset shows an average of 241 release events with incomplete recovery (arrows). **c**, Four single release events (taken from a train of 44 consecutive events) showing variability in release but not in recovery. Red bar indicates stimulation. **d**, Amplitudes of the first consecutive ten stimulations showing that amplitudes varied little from run to run. Error bars indicate s.d., $n = 8$, red line shows line fit. **e**, Average responses to stimuli as in **a**, but with pulse amplitudes of 100 μ A and 200 μ A ($n = 5$). Fluorescence normalized to mean pre-stimulus baseline. For both stimuli, recovery was fitted by a single exponential.

$\tau_{\text{on}} = 134$ ms (100- μ A stimulus) or 62 ms (200- μ A stimulus). In contrast, fluorescence recovery for both stimuli were similar with $\tau_{\text{off}} = 51$ ms (100 μ A) or 61 ms (200 μ A). This shows that vesicle accretion was largely independent of vesicle discharge.

When converting the fluorescence of a release site into an equivalent number of vesicles (see Methods) we found that the number of released vesicles per site was approximately the same (Fig. 4a). At release sites containing 175–1,045 vesicles, a 200- μ A current pulse released around 140 vesicles whereas a 100- μ A pulse released around 80 vesicles from a pool of 220–500 vesicles. This indicates that stimulus amplitude and duration were the determinants of vesicle discharge. Moreover, in sites with low vesicle content, the full releasable vesicle pool was labelled. At the beginning and end of the pulse, a sudden change in the fluorescence was apparent, equivalent to around 40 vesicles (not shown). We ascribe this to the voltage dependence of FM1-43 fluorescence^{21,22}, and our data have been appropriately corrected. The resulting de-staining curve could be fitted with a single exponential function yielding an initial release rate of 3 vesicles ms^{-1} , thereafter slowing exponentially. Fluorescence recovery after the stimulus was equivalent to an initial rate of 1.9 vesicles ms^{-1} and exponential time constant of 87 ms. Averaging over all release events evoked by the 200- μ A stimulus ($n = 241$), we find that the process is described by a release rate of 1.4 vesicles ms^{-1} and a recovery rate of 1.1 vesicles ms^{-1} .

It has long been hypothesized that synaptic ribbons mediate the fast vesicular replenishment required for sustained release during sensory encoding. Ribbon release sites of retinal bipolar cells can refill their releasable pool even in the absence of fast endocytosis, suggesting replenishment from preformed vesicles in the cytoplasm^{23,24}. In IHCs, even long depolarizations cannot deplete the pool of fusion-competent vesicles, suggesting near-inexhaustible pools¹⁸. Our data are compatible with rapid repopulation of IHC

ribbons from pools of preformed vesicles ($\tau < 60$ ms) (Fig. 3e). Paired-pulse experiments had previously narrowed down replenishment time constants to about 140 ms, but the source of vesicles remained unclear³. The high instantaneous release rate (3 vesicles ms^{-1} , Fig. 4c) comes close to the estimated 3–6 vesicles per average excitatory postsynaptic current based on recordings from postsynaptic boutons²⁵, compatible with multi-vesicular release^{25,26} or compound exocytosis^{26,27}. A 150-ms depolarization led to the release of about 170 vesicles (Fig. 4a, c), consistent with previous capacitance-based estimates³. As cytoplasmic vesicles in cells containing ribbon synapses appear to be highly mobile^{28,29}, our data support the idea that the ribbon might serve as a ‘vesicle trap’ capturing cytoplasmic release-ready¹⁸ vesicles after random collision with it^{28,29}, concentrating them at the site of calcium influx. This mechanism would extend the effective releasable pool^{12,13,18} and help ribbon synapses to overcome the constraints of the endocytic vesicle cycle, which keeps maximal synaptic output at 10 vesicles s^{-1} in conventional synapses^{6,7,30}. □

Methods

Tissue preparation, labelling protocol and solutions

Adult guinea pigs (300–500 g) were killed by rapid cervical dislocation according to UK Home Office animal care guidelines. As previously described¹⁴, the organ of Corti remains undissected within the temporal bone so that only apical but not basolateral epithelial surfaces are accessible to the bath solution, which consisted of (in mM): NaCl 144, HEPES 4.9, glucose 23, CaCl_2 1.5, MgCl_2 1.5. Cytoplasmic vesicle pools were labelled with the fluorescent membrane marker FM1-43, bath-applied (5 μM) for 5 min. FM1-43 was internalized exclusively from apical membranes by endocytosis¹⁴. The process leads to labelling of vesicles in cytoplasm and at release sites¹⁴. The rate of apical endocytosis was set to its maximum in all experiments by elevating bath Ca^{2+} to 1.5 mM¹⁴.

Electrical stimulation

Synaptic release was triggered by transepithelial stimulation^{14,15}. Current pulses were passed to an indifferent earth through a 10- μm diameter pipette containing bath solution and placed 40 μm above the IHC stereocilia. The stimuli consisted either of trains of 20-ms-long pulses or, for line-scan imaging, of single pulses of 150-ms duration. Trains typically consisted of 750 pulses delivered at either 30 or 40 Hz, resulting in train durations of 18.8 and 25 s, respectively. Experiments with calcium antagonists were analysed using several imaging planes and 50-s trains. Both pulse frequency and current amplitude were varied. A 200- μA pulse was found to elicit maximal release without stimulation artefacts. In line-scan experiments, current amplitudes were either 100 or 200 μA .

Imaging

Imaging was performed at 22 °C using a two-photon confocal laser scanning microscope (BioRad) consisting of a MRC 1024 scan head mounted on a Nikon microscope with a $\times 60$ 1.0 NA dipping objective. FM1-43 was excited by a Tsunami Ti-Sapphire laser pumped by a Millennia V green laser (Spectra Physics) tuned to 835 nm. Emitted fluorescence was captured by external detectors. The optical slice was 300–500 nm thick. Images (512 $\times$ 512 voxels) were acquired at 0.9 Hz. In most experiments three to four IHCs were imaged simultaneously. Z-stack time series of single IHCs were acquired before and after stimulation to ensure that there was no tissue movement. Reconstructed volumes extending ± 2 μm around release sites were acquired and their integrated intensity has been given in the figures. For high temporal resolution (2 ms) experiments, line scans were used. Individual lines were positioned to section a vesicle cluster. When the line passed near the plasma membrane, a signal from FM1-43 acting as a voltage sensor could be detected^{21,22}. Although representing the presence of FM1-43 in the membrane due to vesicular release, this small offset was subtracted. High-threshold release sites included in line scans served as controls for movement, and thus only scans containing both low- and high-threshold release sites were included in the analysis.

Analysis

Images were analysed using Metamorph (Universal Imaging Corp.), Lucida (Kinetic Imaging) or in-house software. Background fluorescence from areas outside IHCs was subtracted. Photobleaching was $< 1\%$ per min and generally negligible. FM1-43 fluorescence was converted into equivalent 30-nm-diameter vesicle number as described previously by identifying the fluorescence associated with an element of hair cell plasma membrane¹⁴. In line-scan experiments, the number of released vesicles per release site was calculated based on the loss of vesicles within the line voxel volume, assuming that de-staining of synaptic bodies occurred uniformly. A subset of release sites ($n = 7$) was selected for kinetic analysis of vesicular release and replenishment. These release sites showed particularly low baseline noise; that is, the rapid component exceeded one standard deviation (s.d.) of the baseline fluorescent signal. Unless otherwise stated, data are given as mean $\pm$ s.d.

Electrophysiology

For electrophysiological experiments, adult IHCs were recorded at room temperature in strips of organ of Corti perfused with artificial perilymph consisting of (in mM): NaCl 144,

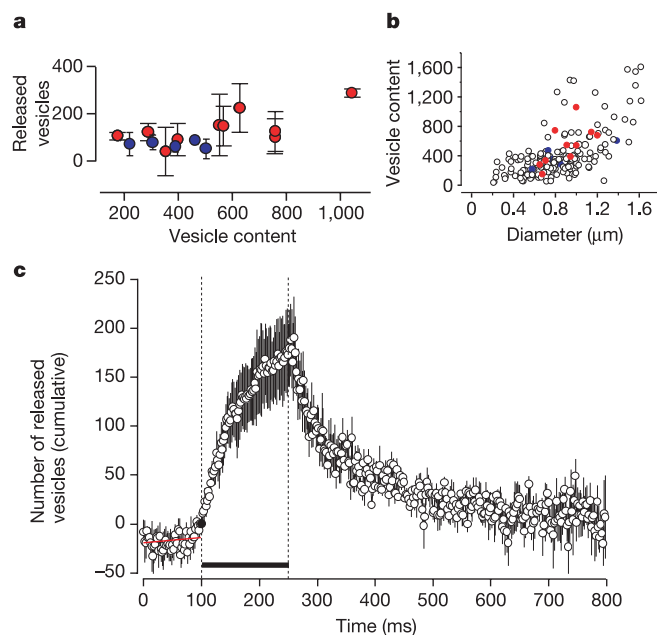


Figure 4 Quantification of vesicular release and recovery. **a**, Peak number of released vesicles as a function of vesicle content before release. Data from 15 release sites. Red circles, 200- μ A stimulus pulse amplitude; blue circles, 100- μ A stimulus pulse amplitude; pulse duration 150 ms. Each data point is the average of 24 release events. Error bars indicate s.d. **b**, The release sites studied with line scans and pulses of 100 μA (blue) and 200 μA (red) were a representative sample of the total release site population analysed regarding both diameter and vesicle content. **c**, Quantification of vesicular release and replenishment for seven release sites. The curve is the average of 24 runs at each of 7 release sites. Error bars indicate s.e.m.

Retinoic acid coordinates somitogenesis and left–right patterning in vertebrate embryos

Julien Vermot & Olivier Pourquié

Stowers Institute for Medical Research, 1000E 50th Street, Kansas City, Missouri 64110, USA

A striking feature of the body plan of a majority of animals is bilateral symmetry. Almost nothing is known about the mechanisms controlling the symmetrical arrangement of the left and right body sides during development. Here we report that blocking the production of retinoic acid (RA) in chicken embryos leads to a desynchronization of somite formation between the two embryonic sides, demonstrated by a shortened left segmented region. This defect is linked to a loss of coordination of the segmentation clock oscillations¹. The lateralization of this defect led us to investigate the relation between somitogenesis and the left–right asymmetry machinery^{2,3} in RA-deficient embryos. Reversal of the situs in chick^{4,5} or mouse⁶ embryos lacking RA results in a reversal of the somitogenesis laterality defect. Our data indicate that RA is important in buffering the lateralizing influence of the left–right machinery, thus permitting synchronization of the development of the two embryonic sides.

In the vertebrate embryo, bilateral symmetry is particularly obvious at the level of the arrangement of somites, which form two symmetrical columns of mesodermal segments on both sides of the embryonic axis. Somites are epithelial spheres generated in a rhythmic fashion from the mesenchymal presomitic mesoderm (PSM), and they subsequently differentiate to give rise to the vertebrae and skeletal muscles of the body. Somite formation involves a molecular oscillator—the segmentation clock—that controls the rhythmic transcription in the PSM of a group of genes called cyclic genes¹. The periodic signal of the oscillator is converted into the metameric array of somite boundaries by a spacing mechanism relying on a travelling threshold of signalling by fibroblast growth factor (FGF) and Wnt, which defines a determination front regressing in concert with body axis extension⁷. The position of the determination front was recently shown to be further refined by a gradient of RA originating from the segmented region and antagonizing the FGF signalling gradient^{8,9}. Both the oscillations of the segmentation clock and the regression of the determination front are synchronous between the two embryonic sides, resulting in the simultaneous production of somites from the PSM on the right and the left side.

We investigated the effect of RA deprivation on somitogenesis in cultured chick embryos by inhibiting RA synthesis with disulphiram, an inhibitor of the RA biosynthetic enzyme retinaldehyde dehydrogenase 2 (RALDH2)¹⁰. To confirm first that treatment with disulphiram blocks RA production, we measured the levels of RA by reverse-phase high-performance liquid chromatography (RP-HPLC) in treated and non-treated embryos¹¹. In seven of eight experiments, little or no RA was detected in pools of treated embryos, whereas RA was detected in nine pools of control embryos (Fig. 1a). Because blocking RA signalling in *Xenopus*, quail and mouse results in an anterior extension of the *fgf8* mRNA gradient in the PSM^{8,9,12}, we examined the *fgf8* expression domain in disulphiram-treated embryos (Fig. 1b, c). In 8-somite to 15-somite control embryos, the rostral boundary of the *fgf8* mRNA gradient was seen at the level of somite –IV (ref. 13) in the caudal PSM ($n = 9$; Fig. 1b). In contrast, disulphiram-treated embryos observed at the same stages showed a rostral expansion of the *fgf8* expression domain up to the level of somite –I (9 of 17; Fig. 1c). Treatment of

KCl 4.6, HEPES 4.9, glucose 23, CaCl₂ 1.5, MgCl₂ 1.5, NaHPO₄ 0.7. Conventional patch-clamp methods were used, with pipette solutions (in mM): Cs gluconate 144, MgCl₂ 2, EGTA 2, TEA 13, HEPES 10, D-glucose 3, Na₂ATP 4, Na⁺-phosphocreatine 10, creatine phosphokinase 20 U ml^{–1}; pH 7.2, osmolality 310 mosmol l^{–1}. In some experiments Ca²⁺ was raised to 8 mM. These higher concentrations accelerated run down and therefore Ba²⁺ ions (at 8 or 20 mM) largely replaced Ca²⁺ (remaining concentration of 0.1 mM). Na⁺ and other ions were adjusted accordingly to maintain osmotic balance. All drugs were bath applied: L-type blockers (nimodipine and nifedipine) at 10 μM, ω-conotoxin GVIA at 1–3 μM, Bay K 8644 at 1 μM.

Received 12 January; accepted 24 March 2005; doi:10.1038/nature03567.

Published online 13 April 2005.

- Liberman, M. C. Efferent synapses in the inner hair cell area of the cat cochlea: an electron microscopic study of serial sections. *Hear. Res.* **3**, 189–204 (1980).
- Saito, K. Fine structure of the sensory epithelium of the guinea pig organ of Corti: afferent and efferent synapses of hair cells. *J. Ultrastruct. Res.* **71**, 222–232 (1980).
- Moser, T. & Beutner, D. Kinetics of exocytosis and endocytosis at the cochlear inner hair cell afferent synapse of the mouse. *Proc. Natl Acad. Sci. USA* **97**, 883–888 (2000).
- Schikorski, T. & Stevens, C. F. Quantitative ultrastructural analysis of hippocampal excitatory synapses. *J. Neurosci.* **17**, 5858–5867 (1997).
- Stevens, C. F. & Wang, Y. Facilitation and depression at single central synapses. *Neuron* **14**, 795–802 (1995).
- Rosenmund, C. & Stevens, C. F. Definition of the readily releasable pool of vesicles at hippocampal synapses. *Neuron* **16**, 1197–1207 (1996).
- Fernandez-Alfonso, T. & Ryan, T. A. The kinetics of synaptic vesicle pool depletion at CNS synaptic terminals. *Neuron* **41**, 943–953 (2004).
- Kiang, Y. Stimulus coding in the auditory nerve and cochlear nucleus. *Acta Otolaryngol. (Stockh.)* **59**, 186–200 (1965).
- Johnson, D. H. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J. Acoust. Soc. Am.* **68**, 1115–1122 (1980).
- Palmer, A. R. & Russell, I. J. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hear. Res.* **24**, 1–15 (1986).
- Sewell, W. F. in *The Cochlea, Springer Handbook of Auditory Research* Vol. 8 (eds Dallos, P., Popper, A. N. & Fay, R. R.) Ch. 9, 503–534 (Springer, New York, 1996).
- Eisen, M. D., Spassova, M. & Parsons, T. D. Large releasable pool of synaptic vesicles in chick cochlear hair cells. *J. Neurophysiol.* **91**, 2422–2428 (2004).
- Edmonds, B. W., Gregory, F. D. & Schweizer, F. E. Evidence that fast exocytosis can be predominantly mediated by vesicles not docked at active zones in frog saccular hair cells. *J. Physiol. (Lond.)* **560**, 439–450 (2004).
- Griesinger, C. B., Richards, C. D. & Ashmore, J. F. FM1-43 reveals membrane recycling in adult inner hair cells of the mammalian cochlea. *J. Neurosci.* **22**, 3939–3952 (2002).
- Mammiano, F. & Ashmore, J. F. Reverse transduction measured in the isolated cochlea by laser Michelson interferometry. *Nature* **365**, 838–841 (1993).
- von Gersdorff, H., Vardi, E., Matthews, G. & Sterling, P. Evidence that vesicles on the synaptic ribbon of bipolar neurons can be rapidly released. *Neuron* **16**, 1221–1227 (1996).
- Spicer, S. S., Thomopoulos, G. N. & Schulte, B. A. Novel membranous structures in apical and basal compartments of inner hair cells. *J. Comp. Neurol.* **409**, 424–437 (1999).
- Beutner, D., Voets, T., Neher, E. & Moser, T. Calcium dependence of exocytosis and endocytosis at the cochlear inner hair cell afferent synapse. *Neuron* **29**, 681–690 (2001).
- Siegel, J. H. & Brownell, W. E. Presynaptic bodies in outer hair cells of the chinchilla organ of Corti. *Brain Res.* **220**, 188–193 (1981).
- Slepecky, N. B., Galsky, M. D., Swartzentruber-Martin, H. & Savage, J. Study of afferent nerve terminals and fibers in the gerbil cochlea: distribution by size. *Hear. Res.* **144**, 124–134 (2000).
- Smith, C. B. & Betz, W. J. Simultaneous independent measurement of endocytosis and exocytosis. *Nature* **380**, 531–534 (1996).
- Neves, G. & Lagnado, L. The kinetics of exocytosis and endocytosis in the synaptic terminal of goldfish retinal bipolar cells. *J. Physiol. (Lond.)* **515**, 181–202 (1999).
- von Gersdorff, H. & Matthews, G. Depletion and replenishment of vesicle pools at a ribbon-type synaptic terminal. *J. Neurosci.* **17**, 1919–1927 (1997).
- Heidelberger, R., Sterling, P. & Matthews, G. Roles of ATP in depletion and replenishment of the releasable pool of synaptic vesicles. *J. Neurophysiol.* **88**, 98–106 (2002).
- Glowatzki, E. & Fuchs, P. A. Transmitter release at the hair cell ribbon synapse. *Nature Neurosci.* **5**, 147–154 (2002).
- Lenzi, D., Crum, J., Ellisman, M. H. & Roberts, W. M. Depolarization redistributes synaptic membrane and creates a gradient of vesicles on the synaptic body at a ribbon synapse. *Neuron* **36**, 649–659 (2002).
- Parsons, T. D. & Sterling, P. Synaptic ribbon. Conveyor belt or safety belt? *Neuron* **37**, 379–382 (2003).
- Zenisek, D., Steyer, J. A. & Almers, W. Transport, capture and exocytosis of single synaptic vesicles at active zones. *Nature* **406**, 849–854 (2000).
- Holt, M., Cooke, A., Neef, A. & Lagnado, L. High mobility of vesicles supports continuous exocytosis at a ribbon synapse. *Curr. Biol.* **14**, 173–183 (2004).
- Stevens, C. F. & Tsujimoto, T. Estimates for the pool size of releasable quanta at a single central synapse and for the time required to refill the pool. *Proc. Natl Acad. Sci. USA* **92**, 846–849 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the MRC, the Wellcome Trust and a BBSRC Imaging grant to C.D.R. We thank U. Egert for his help with part of the image analysis.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.B.G. (claudius.griesinger@physiologie.uni-freiburg.de) or J.F.A. (j.ashmore@ucl.ac.uk).

cultured embryos with disulphiram caused an increase in size of the PSM (28 of 41; compare the distance between arrowheads in Fig. 2a and Fig. 2c), as well as a typical somitic compaction phenotype (compare the last-formed somites in Fig. 2f and Fig. 2g) also seen in vitamin A-deficient quails¹⁴ and *Raldh2* mouse mutant embryos¹⁵. Together, these results indicate that treatment with disulphiram reliably blocks RA signalling in the developing chick embryo.

In addition to the defects in somite size described above, we observed left–right somitic misalignments in the disulphiram-treated embryos. Treated embryos displayed either a delay in the formation of one somite (3 of 41; compare Fig. 2b with Fig. 2d), or had their somite boundaries shifted anteriorly on the left side (16 of 41; Fig. 2g). This desynchronization phenotype could be rescued by adding RA to the disulphiram-treated embryos (15 asymmetric embryos of 24 treated with disulphiram compared with 5 asymmetric embryos of 20 treated with disulphiram and RA; see Methods and Supplementary Fig. 2). These results therefore indicate that in chick embryo, RA signalling might be required for the left–right coordination of somitogenesis.

We next investigated the molecular basis of the asymmetry in somite development observed in RA-deficient embryos. According to current models of somitogenesis, such as the clock and wave-

front, the positioning of somitic boundaries depends on the period of the segmentation clock oscillation and on the speed of the determination front regression⁷. Asymmetric modification of any of these two parameters could therefore result in a change in somitic boundary position, such as that observed in the disulphiram-treated embryos. In disulphiram-treated embryos, no clear asymmetry of *fgf8* expression in the PSM could be observed (Fig. 1c), indicating that asymmetric positioning of the determination front might not be responsible for the somite desynchronization phenotype. We then examined whether disulphiram treatment alters the coordination of segmentation clock oscillations by monitoring the expression of the cyclic gene *Lunatic fringe* (*Lfng*) in the PSM. We analysed waves of *Lfng* expression in disulphiram-treated embryos ranging from stage HH7 to 15 somites. We observed that *Lfng* expression in the PSM was symmetrical in control embryos ($n = 23$; Fig. 1d); whereas, in a majority of treated embryos, *Lfng* expression was asymmetrical (14 of 22). Interestingly, in most treated embryos, the *Lfng* expression domain was found to extend more anteriorly on the left than on the right embryonic side (Fig. 1e). Together, these observations reveal that oscillations of the segmentation clock are accelerated on the left embryonic side in comparison with the right embryonic side. Thus, the primary target of the uncoordination process is the oscillations. These observations are consistent with

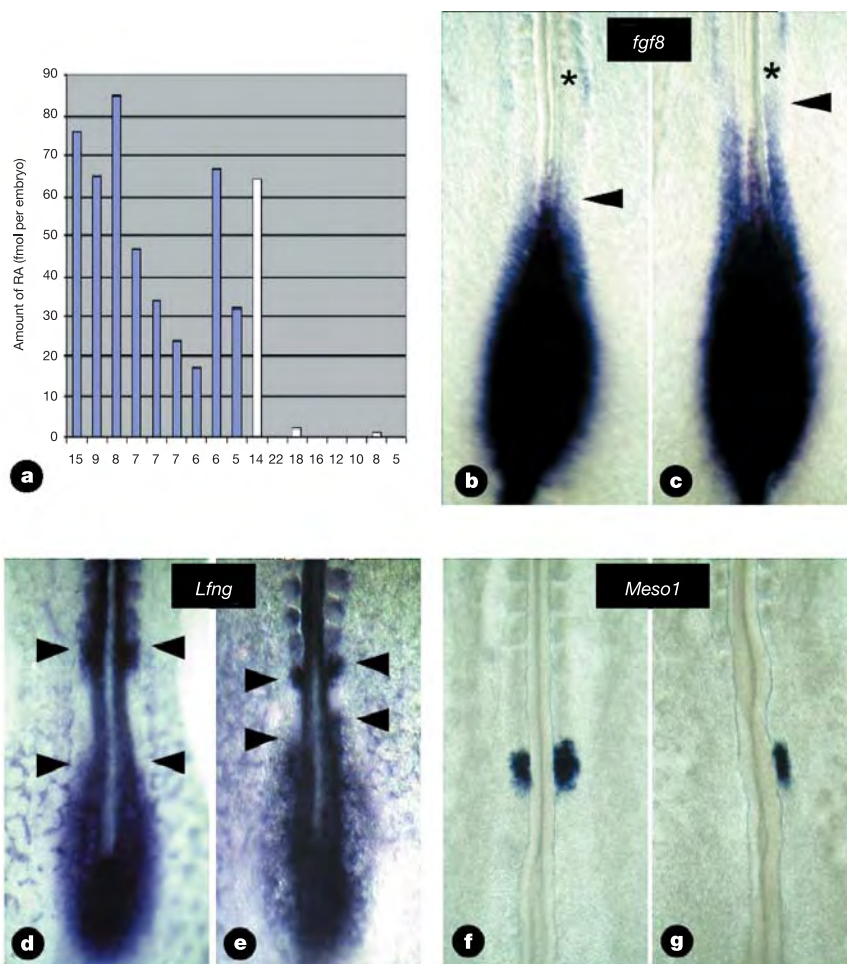


Figure 1 Disulphiram blocks RA production and alters left–right coordination of the segmentation clock. **a**, Comparison of the amount of RA per embryo detected by RP–HPLC in pools of control (blue bars, $n = 9$) and disulphiram-treated (white bars, $n = 8$) chick embryos. The numbers on the x axis indicate the number of embryos in each pool. **b, c**, *In situ* hybridization for *fgf8* in control (**b**) and disulphiram-treated (**c**) embryos.

The arrowhead marks the anterior limit of the *fgf8* domain. The asterisk marks the last-formed somite. **d, e**, *Lfng* expression in control (**d**) and disulphiram-treated (**e**) embryos. Arrowheads mark the anterior limit of *Lfng* expression domains. **f, g**, *Meso1* expression in control (**f**) and disulphiram-treated (**g**) embryos. All views are ventral, with anterior at the top.

previous studies showing that *Lfng* is required for the coordination of somite formation¹⁶.

To test whether the asymmetrical expression of *Lfng* is followed by a desynchronization of the segmentation cascade leading to somite formation, we used *Meso1* as a marker for segmental determination¹⁷. *Meso1* and *Meso2* are expressed in an on/off fashion at the determination front level, indicating that like their mouse homologue *Mesp2*, they might mark the initial segmental response to the periodic clock signal. Half of the control embryos did not express *Meso1*, whereas the other half expressed the gene in a stripe of cells at the level of somite –II ($n = 16$; Fig. 1f). In

disulphiram-treated embryos in which *Meso1* was detected, 3 of 11 embryos displayed an uncoordinated expression pattern, with one band of expression on one side and no expression on the other side (Fig. 1g). Together, these observations indicate that the desynchronization of the segmentation clock oscillations might be responsible for the uncoordination of the somitogenesis process in embryos lacking RA signalling.

If RA deprivation were simply disrupting left–right coordination, random somite desynchronization would be expected. However, our observations show that the segmented region is consistently shortened on the left side, causing somite formation to seem delayed on the left side (Fig. 2e). We also noted that asymmetries arise more often at the level of somite 7–8, which corresponds to the beginning of the cervical region (Fig. 2h). Most of the asymmetries were induced when the disulphiram treatment was performed between stages HH4 and HH6 (8 asymmetrical embryos of 21, and 12 of 24, respectively), soon after the production of the precursors of the anteriormost cervical region by the primitive streak (Fig. 2i)¹⁸. Efficiency decreased at later stages when more posterior somites were produced (between HH7 and HH9, 4 asymmetrical embryos of 29; Fig. 2i). Strikingly, the stages at which disulphiram treatment is most efficient in inducing asymmetric segmentation correspond to the stages at which left-sidedness is transferred from the node to the left lateral plate mesoderm². The left–right bias that we observed indicates that our RA deprivation experiments might reveal an effect of the left–right signalling pathway on somitogenesis. To test this possibility we analysed the influence of the situs on the polarity of the somitogenesis defect. Among the different factors able to alter situs determination in chick embryos, we decided to use Sonic Hedgehog (SHH) because it does not interfere with somite segmentation^{4,5}. SHH-expressing beads were grafted on the right side of the node at stage HH4 to randomize situs determination in RA-deficient chick embryos⁴. In control embryos, SHH beads induced right-sided expression of *Pitx2* in eight of ten embryos, whereas only left-sided *Pitx2* was observed when control beads were grafted (data not shown). In both cases, somitogenesis proceeded symmetrically as in non-grafted embryos. When disulphiram-treated embryos were grafted with a control bead, the left-sided bias in somitogenesis was similar to that described for the non-grafted treated embryos described above (eight left-biased embryos of ten that were asymmetric; Fig. 3a, c–e). The ratio of left-sided embryos changed markedly after the graft of a SHH bead (9 right-biased embryos of 21 that were asymmetric; Fig. 3b, c, f, g). Randomization of the situs in the SHH-bead-grafted embryos was confirmed by analysing *Pitx2* expression (35 embryos of 49 expressed *Pitx2* bilaterally; compare Fig. 3d–f with Fig. 3l–n). When disulphiram-treated embryos displaying bilateral expression of *Pitx2* after the graft of a SHH bead were analysed with *Lfng* as a marker of the oscillations ($n = 13$), embryos showing a right-sided shift of somitogenesis ($n = 3$) displayed a mirror-image expression pattern of *Lfng* compared with embryos exhibiting a left-sided shift of somitogenesis ($n = 3$; compare Fig. 3d, e with Fig. 3f, g). The same mirror-image expression pattern correlating with *Pitx2* expression was observed with *Meso2* for three of five right-shifted embryos (compare Fig. 3h, i with Fig. 3j, k, and Fig. 3l with Fig. 3m); in the other two right-shifted embryos, *Meso2* was expressed on only one side (Fig. 3n). Taken together, these results demonstrate that the asymmetric segmentation defect observed in RA-deficient embryos is under the control of the left–right signalling pathway.

Somite desynchronization defects, resulting in an impaired somitogenesis on the right side comparable to those reported here in disulphiram-treated chick embryos, have been observed in mouse *Raldh2*-null mutants¹². To determine whether these defects are also caused by interference with the left–right machinery, we used the mouse *iv* mutation, which is known to randomize situs determination, to generate embryos that are double mutant for *Raldh2* and *iv* (ref. 6). The double mutants *Raldh2*^{−/−}*iv*^{−/−}

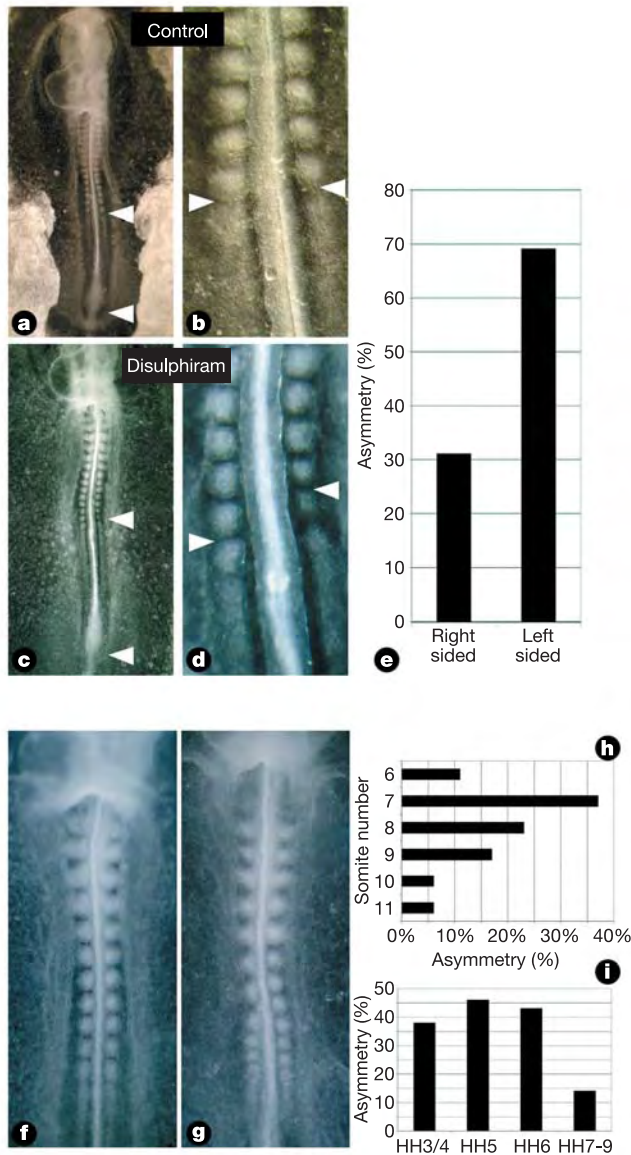


Figure 2 Asymmetry of somitogenesis in disulphiram-treated embryos. **a–d, f, g**, Somite morphology in control (**a, b, f**) and disulphiram-treated (**c, d, g**) embryos. Arrowheads in **a** and **c** mark the antero-posterior limits of the PSM. **e**, Proportions of left-sided and right-sided shifts of somite formation in disulphiram-treated embryos. **f, g**, Embryo displaying a typical left rostral shift of somitogenesis as well as a somitic compaction phenotype of the last formed somites (**g**) compared with a stage-matched control (**f**). **h**, Distribution of somitic misalignments along the anterior–posterior axis. **i**, Effect of age at treatment on the percentage of asymmetry. All views are ventral, with anterior at the top.

displayed all the combinations of *Pitx2* expression in the lateral plate mesoderm: no expression (9 of 29), bilateral expression (9 of 29), right-sided (5 of 29) and left-sided (6 of 29) (data not shown). This distribution is similar to that in single *iv* mutants, confirming that the left–right machinery is functional in *Raldh2*-null embryos^{12,19}. To examine the somitic defects in the mutants, we labelled embryos with *Uncx4.1*, which is expressed in the posterior part of the somites together with *Pitx2* to display the situs. In *Raldh2*^{-/-} and in *Raldh2*^{-/-}*iv*^{+/-} asymmetrical embryos, the right embryonic side always displayed fewer somites than the left side (Fig. 4a, b). In a fraction of *Raldh2*^{-/-}*iv*^{+/-}, alteration in situs determination was accompanied with an inversion of the side of the somitogenesis asymmetry, with the shorter segmented region now on the left side (in six embryos showing either bilateral or no *Pitx2* expression, and in two embryos showing right-sided expression; Fig. 4a, c). Inversion of the somitogenesis defect was also correlated with an inversion of the *Lfng* asymmetrical pattern in the PSM, which was found to be more advanced in its phase of expression on

the left side in double mutants showing an inverted or randomized situs as demonstrated by *Pitx2* labelling (*n* = 6), whereas it was more advanced in the right side in the *Raldh2* mutant (*n* = 14; Fig. 4d, e). These results therefore indicate that in chick and mouse embryos the asymmetric somitogenesis defect uncovered by RA deprivation is controlled by the left–right signalling machinery. Interestingly, the somitogenesis defect is found on opposite sides in chick and mouse embryos. This discrepancy is consistent with the different function of FGF8, which acts as a right determinant in chick and as a left one in the mouse²⁰.

Recent evidence shows that *Nodal* activation in the node requires Notch signalling^{21,22}. In chick embryos, an initial asymmetry in membrane voltage potential set up across the primitive streak by a gradient of H⁺/K⁺-ATPase activity establishes a left–right extracellular gradient of Ca²⁺. This gradient triggers Notch hyperactivation on the left side of Hensen's node by a mechanism temporally controlled by the waves of *Lfng* that periodically sweep the paraxial mesoderm around the node. This mechanism, acting together or in

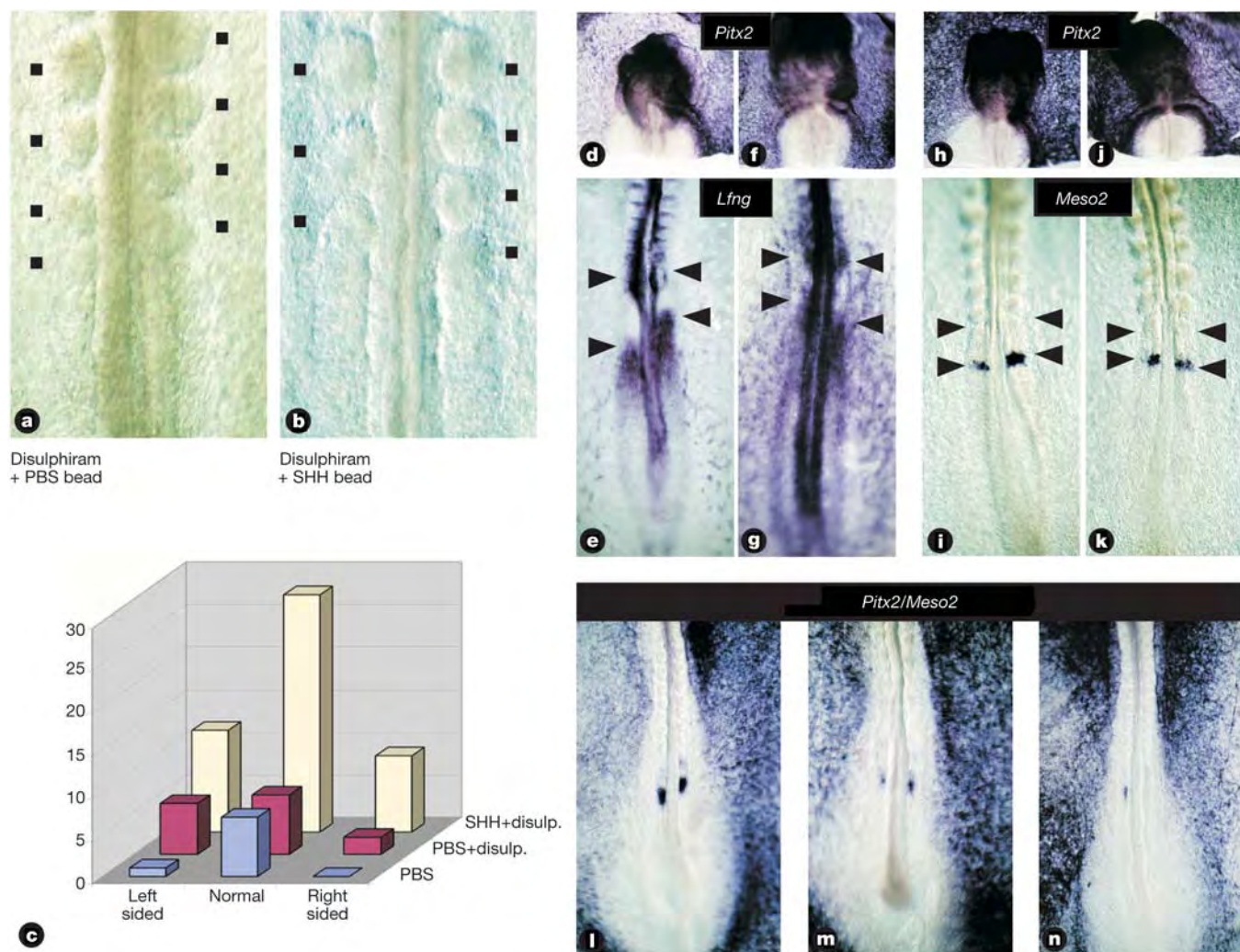


Figure 3 Inversion of the situs leads to an inversion of the somitic defect in embryos lacking RA signalling. **a, b**, Somitic morphology after disulphiram treatment in embryos grafted with control beads (**a**) and SHH beads (**b**), showing a mirror image of the somite misalignment (dots label the position of somites). **c**, Numbers of embryos harbouring left-sided and right-sided somitogenesis defects after grafting of control or SHH beads in disulphiram-treated embryos or in controls. **d–k**, *In situ* hybridization with *Pitx2* (**d, f, h, j**) of the anterior region of the embryos whose posterior region is shown in the panels below, after hybridization with *Lfng* (**e, g**) and *Meso2* (**i, k**) after implantation with control beads

(**d, e, h, i**) or SHH beads (**f, g, j, k**). In embryos grafted with a control bead and treated with disulphiram, *Pitx2* expression was left-sided (**d, h**). In contrast, *Pitx2* expression was bilateral after SHH bead implantation in disulphiram-treated embryos (**f, j**). **l**, Disulphiram-treated embryo grafted with a control bead showing a left-sided expression of *Pitx2* and a rostral shift of *Meso2* expression on the left. **m, n**, Disulphiram-treated embryo grafted with a SHH bead showing a bilateral expression of *Pitx2* and a rostral shift of *Meso2* on the right (**m**) or *Meso2* expression on only one side (**n**). All views are ventral, with anterior at the top.

parallel with the SHH pathway, results in the left-sided expression of *Nodal* in the early organizer^{3,23}. Because *Lfng* is also a downstream target of Notch signalling¹⁶, a consequence of the asymmetric Notch signalling around Hensen's node is the progressive desynchronization of the waves of *Lfng* expression²³. Thus, *in vivo*, initiating the left-right asymmetry pathway leads to the transient desynchronization of the segmentation clock. However, no somitic asymmetries reflecting this early desynchrony of the segmentation clock were observed in embryos, indicating that it might subsequently be compensated for through a different mechanism. Our results show that in chick and mouse embryos, RA signalling is important in compensating for this asymmetrical influence, coming from the left-right signalling pathway on the coordination of somitogenesis. Parallel observations in the zebrafish embryo²⁴ indicate that the function of RA in controlling somitogenesis symmetry is conserved in vertebrates. However, in more primitive chordates such as

Amphioxus, the somites are not symmetrical²⁵, indicating that this function for RA might be specific to vertebrates. Whether RA has a broader function in the control of left-right symmetry in vertebrate embryos remains to be explored. □

Methods

Chicken embryo manipulation and *in situ* hybridization

Fertilized eggs were obtained from Ozark Hatcheries (Neosho, Missouri) and incubated at 38 °C. Embryos were staged as described in ref. 26. Embryos were explanted in New culture using the EC method²⁷. For disulphiram treatments, embryos were pretreated for 6 h in chick medium (5% chick serum, 2.5% FCS, 0.18% bicarbonate, in DMEM medium) containing 140 µM disulphiram (Sigma; stock solution 400 mM in dimethylsulphoxide). Embryos were next transferred to agarose-albumin culture plates, and 100 µl of a solution of 800 µM disulphiram in PBS was deposited on the embryos. Somitic alignment was analysed after 20 h of culture at 38 °C. Rescue experiments were performed with the same disulphiram-containing solutions supplemented by 0.7 µM RA (Sigma) in chick medium. Embryos were next treated with 100 µl of a solution containing 800 µM disulphiram and 3.3 µM RA after transfer onto agarose-albumin culture plates. Somitic alignment was scored in the cervical region in 20 rescued embryos and 24 disulphiram-treated embryos. For situs randomization, beads loaded with SHH protein (R&D Systems) were inserted on the right embryonic side at the level of Hensen's node, as described². After 3 h of culture at 38 °C, embryos were treated with disulphiram with the use of the protocol described above. Somite and PSM nomenclatures are as described¹³. Embryos were processed for whole-mount *in toto* hybridization as described²⁸, using probes provided by A. Buchberger (*cMeso1*, *cMeso2*), J. C. Izpisua Belmonte (*cfgf8*), C. J. Tabin (*cPitx2*), P. Gruss (*mUncx4.1*), T. Vogt (*mLfng*) and C. Goriadis (*mPitx2*).

RP-HPLC

RA extraction was performed as described¹¹ from control and disulphiram-treated whole embryos cultured for 20 h at 38 °C. RP-HPLC was performed with a silica column (LiChrospher, 5 µm RP-18; 125 mm × 4 mm; Phenomenex) and System Gold HPLC (Beckman Coulter) on extracts of pools containing between 5 and 22 embryos of disulphiram-treated ($n = 8$) or non-treated ($n = 9$) embryos (Fig. 1a). The column flow-through was analysed by spectrophotometry (A_{351}), and peak values were obtained with the 32 Karat program (Beckman). RA was identified by its retention time on the column (22 min). A standard curve was generated with increasing concentrations of synthetic RA and used to calculate the concentration of RA in embryo extracts (Supplementary Fig. 1). Measures from two independent experiments performed on different days with two different batches of synthetic RA were used to generate the standard curve. The quantities of RA in pools of control ($n = 9$) and disulphiram-treated ($n = 8$; Fig. 1a) embryos was calculated with the equation obtained from the standard curve ($R^2 = 0.9936$, $y = 17.917x + 36.452$; ExcelX). A Mann-Whitney test was performed to verify that the medians of the treated embryos were less than that of the controls ($p = 0.0016$; Minitab statistical software).

Mouse breeding

Raldh2 mice¹⁵ were crossed with *iv* mice provided by the Jackson Laboratory. Double heterozygous mice were intercrossed to obtain double-mutant embryos and controls. Genotyping was performed as described^{15,29}.

Received 10 December 2004; accepted 24 February 2005; doi:10.1038/nature03488.

1. Pourquie, O. The segmentation clock: converting embryonic time into spatial pattern. *Science* **301**, 328–330 (2003).
2. Raya, A. & Belmonte, J. C. Sequential transfer of left-right information during vertebrate embryo development. *Curr. Opin. Genet. Dev.* **14**, 575–581 (2004).
3. Levin, M. Left-right asymmetry in embryonic development: a comprehensive review. *Mech. Dev.* **122**, 3–25 (2005).
4. Levin, M. *et al.* Left/right patterning signals and the independent regulation of different aspects of situs in the chick embryo. *Dev. Biol.* **189**, 57–67 (1997).
5. Pagan-Westphal, S. M. & Tabin, C. J. The transfer of left-right positional information during chick embryogenesis. *Cell* **93**, 25–35 (1998).
6. Brueckner, M., D'Eustachio, P. & Horwich, A. L. Linkage mapping of a mouse gene, *iv*, that controls left-right asymmetry of the heart and viscera. *Proc. Natl Acad. Sci. USA* **86**, 5035–5038 (1989).
7. Dubrulle, J. & Pourquie, O. Coupling segmentation to axis formation. *Development* **131**, 5783–5793 (2004).
8. Diez del Corral, R. *et al.* Opposing FGF and retinoid pathways control ventral neural patterning, neuronal differentiation, and segmentation during body axis extension. *Neuron* **40**, 65–79 (2003).
9. Moreno, T. A. & Kintner, C. Regulation of segmental patterning by retinoic acid signaling during *Xenopus* somitogenesis. *Dev. Cell* **6**, 205–218 (2004).
10. Stratford, T., Horton, C. & Maden, M. Retinoic acid is required for the initiation of outgrowth in the chick limb bud. *Curr. Biol.* **6**, 1124–1133 (1996).
11. Maden, M., Sonneveld, E., van der Saag, P. T. & Gale, E. The distribution of endogenous retinoic acid in the chick embryo: implications for developmental mechanisms. *Development* **125**, 4133–4144 (1998).
12. Vermot, J. *et al.* Retinoic acid controls the bilateral symmetry of somite formation in the mouse embryo. *Science* **308**, 563–566 (2005).
13. Pourquie, O. & Tam, P. P. A nomenclature for prospective somites and phases of cyclic gene expression in the presomitic mesoderm. *Dev. Cell* **1**, 619–620 (2001).
14. Maden, M., Graham, A., Zile, M. & Gale, E. Abnormalities of somite development in the absence of retinoic acid. *Int. J. Dev. Biol.* **44**, 151–159 (2000).
15. Niederreither, K., Subbarayan, V., Dolle, P. & Chambon, P. Embryonic retinoic acid synthesis is essential for early mouse post-implantation development. *Nature Genet.* **21**, 444–448 (1999).

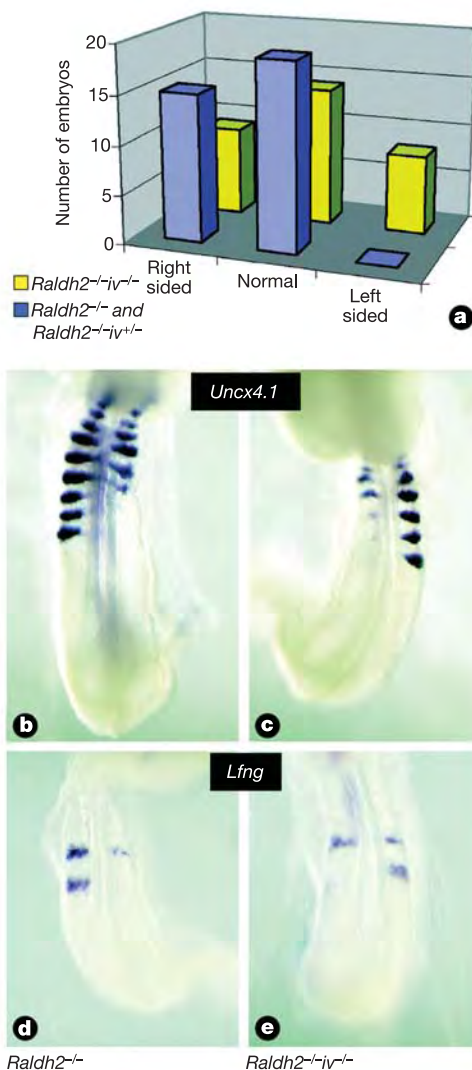


Figure 4 Inversion of somitogenesis asymmetry in *Raldh2*^{-/-} *iv*^{-/-} mouse mutants showing situs inversion. **a**, Numbers of embryos displaying left and right asymmetric somitogenesis scored after *Uncx4.1* mRNA labelling in *Raldh2*^{-/-} and *Raldh2*^{-/-} *iv*^{-/-} (blue bars) and in *Raldh2*^{-/-} *iv*^{-/-} embryos (yellow bars). **b, c**, *In situ* hybridization with *Uncx4.1* in a *Raldh2*^{-/-} embryo (**b**) and in a *Raldh2*^{-/-} *iv*^{-/-} embryo harbouring an inverted phenotype (**c**). **d, e**, *Raldh2*^{-/-} embryo (**d**) and *Raldh2*^{-/-} *iv*^{-/-} embryo (**e**) showing a mirror-image expression pattern of *Lfng*. All views are dorsal, with anterior at the top.

16. Dale, J. K. *et al.* Periodic Notch inhibition by Lunatic Fringe underlies the chick segmentation clock. *Nature* **421**, 275–278 (2003).
17. Buchberger, A., Seidl, K., Klein, C., Eberhardt, H. & Arnold, H. H. cMeso-1, a novel bHLH transcription factor, is involved in somite formation in chicken embryos. *Dev. Biol.* **199**, 201–215 (1998).
18. Psychoyos, D. & Stern, C. D. Fates and migratory routes of primitive streak cells in the chick embryo. *Development* **122**, 1523–1534 (1996).
19. Piedra, M. E., Icardo, J. M., Albajar, M., Rodriguez-Rey, J. C. & Ros, M. A. Pitx2 participates in the late phase of the pathway controlling left-right asymmetry. *Cell* **94**, 319–324 (1998).
20. Meyers, E. N. & Martin, G. R. Differences in left-right axis pathways in mouse and chick: functions of FGF8 and SHH. *Science* **285**, 403–406 (1999).
21. Raya, A. *et al.* Notch activity induces Nodal expression and mediates the establishment of left-right asymmetry in vertebrate embryos. *Genes Dev.* **17**, 1213–1218 (2003).
22. Krebs, L. T. *et al.* Notch signaling regulates left-right asymmetry determination by inducing Nodal expression. *Genes Dev.* **17**, 1207–1212 (2003).
23. Raya, A. *et al.* Notch activity acts as a sensor for extracellular calcium during vertebrate left-right determination. *Nature* **427**, 121–128 (2004).
24. Kawakami, Y., Raya, A., Marina Raya, R., Rodriguez-Esteban, C. & Izpisua Belmonte, J. C. Retinoic acid signalling links left-right asymmetric patterning and bilaterally symmetric somitogenesis in the zebrafish embryo. *Nature* doi:10.1038/nature03512 (this issue).
25. Schubert, M., Holland, L. Z., Stokes, M. D. & Holland, N. D. Three amphioxus Wnt genes (AmphiWnt3, AmphiWnt5, and AmphiWnt6) associated with the tail bud: the evolution of somitogenesis in chordates. *Dev. Biol.* **240**, 262–273 (2001).
26. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).
27. Chapman, S. C., Collignon, J., Schoenwolf, G. C. & Lumsden, A. Improved method for chick whole-embryo culture using a filter paper carrier. *Dev. Dyn.* **220**, 284–289 (2001).
28. Henrique, D. *et al.* Expression of a Delta homologue in prospective neurons in the chick. *Nature* **375**, 787–790 (1995).
29. Adachi, H. *et al.* Determination of left/right asymmetric expression of nodal by a left side-specific enhancer with sequence similarity to a lefty-2 enhancer. *Genes Dev.* **13**, 1589–1600 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Chambon for his interest in this work; P. Dollé for discussions and for providing the *Raldh2* mice; H. Hamada for providing the *iv* genotyping protocol; members of the Pourquie laboratory for sharing reagents and helpful discussions, especially T. Iimura for sharing his expertise with chick culture; B. Brede and P. Malapert for help with mouse genotyping; S. Esteban for artwork; and members of the Conaway and Workman laboratories, particularly J. Conaway and A. Paoletti for help with RP-HPLC, and J. Chen (on leave from University of Missouri – Kansas City) for help with the statistics. J.V. was supported by the Fondation pour la Recherche Médicale and by a Travelling Fellowship from the Company of Biologists and is on leave from IGBMC, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France. Present work is supported by the Stowers Institute for Medical Research and by a grant from the NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to O.P. (olp@stowers-institute.org).

Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice

Maki Nakayama¹, Norio Abiru², Hiroaki Moriyama³, Naru Babaya¹, Edwin Liu¹, Dongmei Miao¹, Liping Yu¹, Dale R. Wegmann¹, John C. Hutton¹, John F. Elliott⁴ & George S. Eisenbarth¹

¹Barbara Davis Center for Childhood Diabetes, University of Colorado Health Sciences Center, Denver, Colorado 80262, USA

²Graduate School of Biomedical Sciences, Nagasaki University, Nagasaki 852-8501, Japan

³Kobe University Graduate School of Medicine, Kobe 650-0017, Japan

⁴University of Alberta, Edmonton, Alberta T6G 2S2, Canada

A fundamental question about the pathogenesis of spontaneous autoimmune diabetes is whether there are primary autoantigens. For type 1 diabetes it is clear that multiple islet molecules are the target of autoimmunity in man and animal models^{1,2}. It is not clear whether any of the target molecules are essential for the

destruction of islet beta cells. Here we show that the proinsulin/insulin molecules have a sequence that is a primary target of the autoimmunity that causes diabetes of the non-obese diabetic (NOD) mouse. We created insulin 1 and insulin 2 gene knockouts combined with a mutated proinsulin transgene (in which residue 16 on the B chain was changed to alanine) in NOD mice. This mutation abrogated the T-cell stimulation of a series of the major insulin autoreactive NOD T-cell clones³. Female mice with only the altered insulin did not develop insulin autoantibodies, insulinitis or autoimmune diabetes, in contrast with mice containing at least one copy of the native insulin gene. We suggest that proinsulin is a primary autoantigen of the NOD mouse, and speculate that organ-restricted autoimmune disorders with marked major histocompatibility complex (MHC) restriction of disease are likely to have specific primary autoantigens.

Mice have two insulin genes, insulin 1 on chromosome 19, and insulin 2 on chromosome 7. Insulin 1 differs from insulin 2 by two amino acids at positions 9 and 29 on the B chain. Breeding a knockout of the insulin 2 gene (produced by J. Jami) into NOD mice results in an accelerated development of diabetes and an enhanced production of insulin autoantibodies^{4,5}. In contrast, similar breeding of the insulin 1 knockout into the NOD mouse prevents most progression to diabetes but does not decrease the expression of insulin autoantibodies; most of these mice develop insulinitis and a smaller subset progress to overt diabetes.

It has recently been reported⁶ that inducing recessive tolerance with a proinsulin 2 construct with an MHC class II invariant chain promoter in NOD mice greatly decreases the development of diabetes. The protection in terms of diabetes was incomplete despite low levels or absence of insulin-reactive T cells after immunization with proinsulin 1 and 2 and low levels or absence of anti-insulin autoantibodies. The authors concluded that insulin is a key but not essential antigen for diabetes of the NOD mouse. They did not use a proinsulin 1 construct; instead they showed cross-reactivity between insulin 1 and insulin 2 by *in vitro* assay. As the authors discussed, small numbers of insulin-reactive T cells that are not detected by their *in vitro* assay might mediate the incomplete prevention of autoimmune diabetes⁶, such as those recognizing insulin 1.

Most CD4 T cells infiltrating NOD islets react to insulin, and more than 90% recognize insulin B chain 9–23 peptide amino acids (insulin B:9–23) (ref. 7). An alanine scan of insulin B:9–23 indicated that changing the native tyrosine to alanine at insulin B chain position 16 (B16) abrogated the response of B:9–23-reactive CD4 T-cell clones³. In addition, it has been reported that a CD8 T-cell clone recognized insulin B:15–23 and the alternative mutation of tyrosine to alanine at B16 results in the failure to bind to the K^d class I MHC molecule⁸. We proposed that both the preproinsulin 1 and 2 B chain 9–23 sequences, differing only at position 9 (serine for insulin 2, proline for insulin 1), would be redundantly important for the development of diabetes in NOD mice and therefore produced NOD mice lacking both native insulin genes. To prevent diabetes in such mice lacking both native insulin genes ('metabolic diabetes') we developed preproinsulin-transgenic strains directly in NOD mice with a mutated sequence. The mutation (alanine rather than tyrosine at B16) was chosen to preserve insulin's metabolic activity but to abrogate T-cell reactivity to B:9–23 (refs 1,3,7,9). The objective of the current study is to determine whether a complete lack of native insulin with B:9–23 sequence would abrogate the development of anti-islet autoimmunity.

Mutated preproinsulin-transgenic mice were produced directly in NOD mice. The transgene expression of two of the founder strains (strains E and H) were unable to prevent the development of diabetes in the absence of any native insulin genes (minimal expression of islet immunoreactive insulin (not shown)), with 'early' (for NOD) development of metabolic diabetes in the absence

of insulinitis (less than 10 weeks of age), whereas two other founder strains prevented this metabolic diabetes (strains B and F) and were used to analyse the development of immune-mediated diabetes in the absence of native insulin sequences.

The NOD transgenic strains were combined with NOD insulin gene knockouts and prospectively evaluated for the production of insulin autoantibodies, insulinitis and diabetes, relative to the presence or absence of the insulin 1 gene. As shown in Fig. 1, female mice lacking both native insulin genes (*ins1*^{-/-} and *ins2*^{-/-}) failed to produce insulin autoantibodies. In contrast, mice with the insulin 1 gene, despite lacking the insulin 2 gene, developed insulin autoantibodies ($P < 0.0001$; peak insulin autoantibodies).

Figure 2 illustrates the histology of native insulin gene double knockout female NOD mice killed at the ages of 26 weeks (Fig. 2a–c; transgene founder strain B) or 23 weeks (Fig. 2d; transgene founder strain F). The pancreas was normal, with insulin-staining cells comprising 0.75% and 1.8%, respectively, of the pancreatic area (standard NOD mice less than 10 weeks of age ($n = 10$) showed an insulin-staining area ranging from 0.74% to 1.56% (median 1.2%)). There was no insulinitis at 26 weeks (Fig. 2a, b) and 23 weeks (Fig. 2d) in the mice lacking native insulin. In contrast, insulinitis was present at 26 weeks (Fig. 2c) and 23 weeks (not shown), indicating that as expected insulin knockouts do not prevent all autoimmunity of NOD mice but do prevent islet-specific autoimmunity. It is extremely unlikely that the insertion of the transgene disrupted a gene important for autoimmunity, given the presence of two founder strains with the same phenotype. In addition, in the presence of the native insulin genes the transgene does not prevent autoimmunity (see below).

Figure 3 illustrates the histology of female mice expressing native insulin that were killed after 10 weeks of age. As expected, all diabetic mice ($n = 15$ of 15; three shown in Fig. 3 representing different genotypes) with at least one insulin 1 or insulin 2 gene and with or without the mutated transgene showed insulinitis. Similarly, almost all non-diabetic mice killed after 20 weeks of age ($n = 25$ of 26, three shown in Fig. 3 representing different genotypes) showed insulinitis ($P < 0.01$ versus double knockouts for insulin). None of the 27 insulin 1⁻ insulin 2⁻ islets had insulinitis, in contrast to mice with native insulin genes (insulin 1⁺ insulin 2⁻: 18 islets with no insulinitis (14%), 22 with peri-islet insulinitis, 88 with intra-islet insulinitis; insulin 1⁻ insulin 2⁺: 99 with no insulinitis (50%), 47 with peri-islet insulinitis, 54 with intra-islet insulinitis; insulin 1⁺ insulin 2⁺: 10 with no insulinitis (23%), 16 with peri-islet insulinitis, 18 with intra-islet insulinitis; $P < 0.001$).

As shown in Fig. 4a, none of the mice lacking the native sequences

of both insulin 1 and insulin 2 developed diabetes, which is consistent with their lack of insulinitis. In contrast, the presence of insulin 1 restored the development of diabetes (75% of mice were diabetic at 25 weeks, by life-table analysis; Fig. 4a).

To exclude the possibility that the prevention of diabetes was related to the chromosomal region from 129 mice of the knockouts bred into the NOD or of the mutated preproinsulin transgene, we analysed the progression to diabetes of multiple control crosses. Figure 4 shows that fixing the control (non-knockout) chromosomal region surrounding the insulin genes (speed congenics backcross no. 6 for the insulin 1 region, backcross no. 4 for the insulin 2 region) on the NOD background did not alter the development of diabetes (Fig. 4b, insulin 1 region; Fig. 4c, insulin 2 region). In addition, all transgenic founder strains with the mutated preproinsulin gene showed progression to diabetes (Fig. 4d) and insulinitis (not shown), with founder strain F having a modest decrease in progression to diabetes and the highest thymic expression of insulin messenger RNA (mRNA data not shown), which might affect autoreactive or regulatory T-cell development with the modified B:9–23 insulin (B16:alanine) or other proinsulin epitopes of the transgene.

To determine whether functional 'diabetogenic' splenocytes could be detected in the insulin knockout mice, we harvested spleen cells from insulin gene double-knockout and wild-type NOD mice and transferred them into immunodeficient (severe combined immunodeficiency; SCID) NOD recipients. Splenocytes were obtained from mice between 9 and 26 weeks of age, a time at which accelerated transfer of diabetes is observed for standard NOD mice. Whereas 75% of recipients were diabetic by 8 weeks after the transfer of wild-type splenocytes, recipients of splenocytes from insulin gene double-knockout NOD mice had a marked delay in developing diabetes (Fig. 4e; $P < 0.02$). These splenocytes from the insulin gene double-knockout NOD mouse induced diabetes 2–4 months after transfer into the NOD.SCID mice. The recipient mice expressed native insulin 1 and insulin 2 in their islets, but this experiment does not prove that normal insulin or insulin B:9–23 is the major target of the transferred cells. The long period to diabetes induction by splenocytes from mice lacking native insulin genes, when transferred into mice with native-insulin-containing islets,

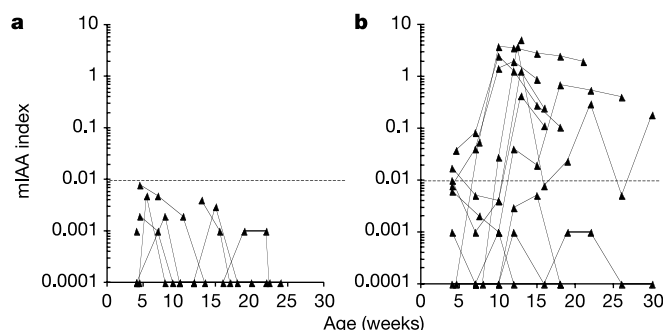


Figure 1 Serum anti-insulin autoantibody levels. **a**, Insulin autoantibodies fail to occur in transgene⁺ mice in the absence of native insulin 1 and insulin 2 genes. **b**, Mice with insulin 1 gene (insulin 1^{+/+}, insulin 2^{-/-}, transgene⁺ or transgene⁻) produce insulin autoantibodies (IAA). Lines connect results for individual mice. mIAA, micro-IAA assay.

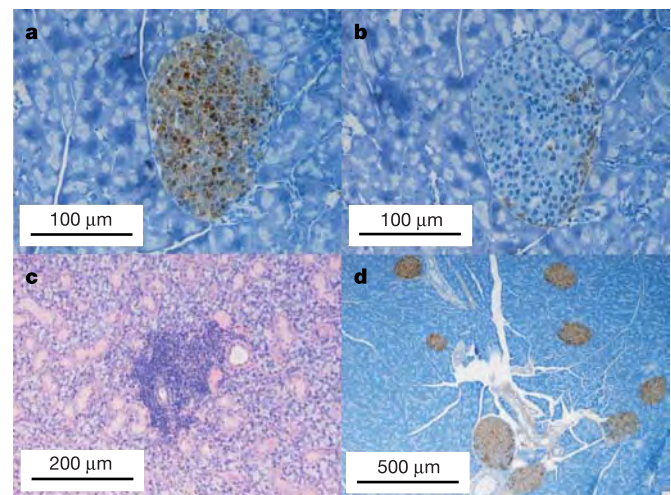


Figure 2 Histology of non-diabetic native insulin-negative mice (insulin 1^{-/-}, insulin 2^{-/-}, transgene⁺). Histology of pancreas and salivary gland of NOD mice lacking both native insulin 1 and 2 genes, with mutated (B16:alanine) preproinsulin transgene. **a–c**, Sections from a mouse killed at 26 weeks, showing islets stained for insulin (**a**), islets stained for glucagon (**b**) and salivary gland stained with haematoxylin and eosin (**c**). **d**, Islets from a mouse killed at 23 weeks, stained for insulin.

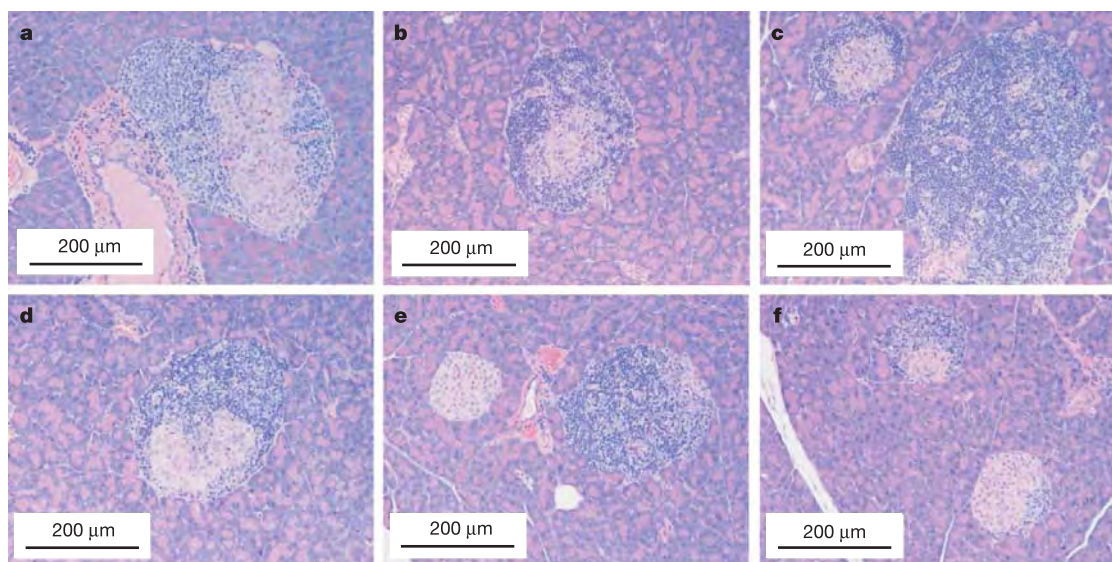


Figure 3 Histology of native insulin-positive mice. Histology showing marked insulitis of NOD mice with one or more copies of the insulin 1 or insulin 2 gene, with or without the mutated preproinsulin (B16:alanine) transgene. **a**, Insulin 1^{+/+}, insulin 2^{-/-}, transgene⁻, killed at 14 weeks, diabetic at 14 weeks. **b**, Insulin 1^{+/+}, insulin 2^{-/-}, transgene⁺, killed at 10.5 weeks, diabetic at 10.5 weeks. **c**, Insulin 1^{+/+}, insulin 2^{-/-}, transgene⁺, killed at 15 weeks, diabetic at 15 weeks. **d**, Insulin 1^{-/-}, insulin 2^{+/+}, transgene⁻, killed at 20 weeks, not diabetic. **e**, Insulin 1^{-/-}, insulin 2^{+/+}, transgene⁺, killed at 21 weeks, not diabetic. **f**, Insulin 1^{-/-}, insulin 2^{+/+}, transgene⁺, killed at 35 weeks, not diabetic.

transgene⁺, killed at 15 weeks, diabetic at 15 weeks. **d**, Insulin 1^{-/-}, insulin 2^{+/+}, transgene⁻, killed at 20 weeks, not diabetic. **e**, Insulin 1^{-/-}, insulin 2^{+/+}, transgene⁺, killed at 21 weeks, not diabetic. **f**, Insulin 1^{-/-}, insulin 2^{+/+}, transgene⁺, killed at 35 weeks, not diabetic.

indicates the possible development of pathogenic T-cell responses after splenocyte transfer.

Type 1A diabetes is characterized by specific destruction of the cells within the pancreas that produce insulin¹⁰, and alleles of class II MHC genes are major determinants of disease in both humans and animal models^{11,12}. Risk of type 1A diabetes varies by more than 1,000-fold depending on the HLA (human leukocyte antigen) DQ genotypes in man. One hypothesis to account for these potent MHC associations and beta-cell-specific destruction is to posit that the immune response for a given genotype (with immune dysregulation

and environmental factors) is dependent on an antigen whose targeting is essential for disease. T-cell clones reacting with multiple antigens of the islets, including novel molecules induced to be expressed within islets, are able to transfer diabetes, providing strong evidence that multiple actual and potential islet targets exist. However, there is evidence that glutamic acid decarboxylase 65 (GAD65), insulinoma antigen 2 (IA-2), IA-2 β /phogrin and heat shock protein 60 (hsp60) are not essential autoantigens of the spontaneous NOD mouse, in that mice develop diabetes despite the removal of the antigen or antigen-responding T cells¹³⁻¹⁷.

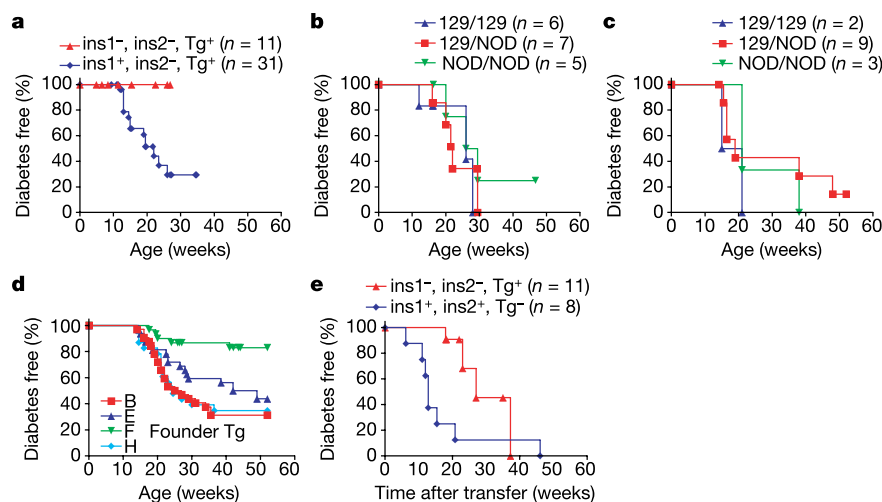


Figure 4 Life tables of diabetes development. **a**, Lack of progression to diabetes of NOD mice lacking both insulin native genes. **b**, **c**, Lack of difference in progression to diabetes for speed congenic female NOD mice with control regions from the 129 mouse surrounding the insulin genes (insulin 1 region (**b**); insulin 2 region (**c**)). **d**, Progression to diabetes of the four transgenic founder strains with the mutated preproinsulin gene on

NOD background (insulin 1^{+/+}, insulin 2^{+/+}, transgene⁺). Founder strain F was significantly different ($P < 0.01$) from the others. **e**, NOD.SCID recipients of splenocytes from mice lacking both native insulin genes had delayed progression to diabetes compared with insulin 1⁺, insulin 2⁺ mice ($P < 0.02$). ins, insulin; Tg, transgene.

Our studies indicate that insulin peptide B:9–23 might be an essential target of the immune destruction of the NOD mouse, although it is possible that the processing of other epitopes in the proinsulin molecule might be affected by the single amino-acid mutation at position B16, and there is the caveat that mutated proinsulin (B16:alanine) transgene expression might not be permissive for islet autoimmunity. There is at present considerable evidence that insulin is an important target molecule, despite evidence for many additional relevant targets^{18–20}. Expression of proinsulin in transgenic lymphoid organs or modulating the immune response to insulin can greatly decrease the development of diabetes^{6,21,22}.

Our studies provide evidence that the development of insulin autoantibodies, insulinitis and diabetes is dependent on native insulin gene sequences. Absence of insulinitis in the presence of sialitis indicates that this influence might be organ-specific. Multiple studies indicate that both insulin 1 and insulin 2, which differ by only 2 of 51 amino acids, can be the target of autoimmunity. If primary autoantigenic targets exist for specific autoimmune disorders with specific MHC genotypes, we believe it likely that deletional therapies may allow potent disease prevention. Such deletional therapies may be more robust than regulatory experimental manipulations, which usually do not abrogate insulinitis^{23,24}. At a minimum, the induction of this deletional tolerance might be a preventive pathway synergistic to the harnessing of regulatory T-cell responses. □

Methods

Mice

Insulin 1 and insulin 2 knockout NOD mice were established by breeding the original insulin knockouts provided by J. Jami into NOD/Bdc mice with the use of speed congenic techniques⁵. The original insulin knockout mice were produced in 129S1/SvImJ embryonal cell lines²⁵. Mutated preproinsulin-transgenic NOD mice were produced by the microinjection of mutated preproinsulin 2 complementary DNA constructs ligated to the pRIP7 (rat insulin 7) promoter directly into fertilized NOD oocytes as described previously²⁶. Female mice were used for the study.

To rule out potential effects of 129S1/SvImJ genes in linkage with the insulin gene, the mice, which have the 129S1/SvImJ gene at insulin 1 (position 49 cM at chromosome 19) or insulin 2 (position 69.1 cM at chromosome 7) locus on the NOD background, were produced with the use of speed congenic methods. We bred 129S1/SvImJ mice with NOD/Bdc mice, and NOD diabetogenic loci (idd 1–14) were fixed by backcross 3. Both strains were further backcrossed into NOD/Bdc mice. 129S1/SvImJ genomic regions flanking each insulin gene locus are less than 18 cM for the insulin 1 region-fixed mice and less than 19 cM for the insulin 2 region mice; the central 129 regions without the knockouts spanned these regions. Female mice were studied.

Immunodeficient NOD mice (NOD.SCID) for the adoptive transfer experiment were purchased from the Jackson Laboratory. All mice were housed in a pathogen-free animal colony at Barbara Davis Center for Childhood Diabetes using an approved protocol from the University of Colorado Health Sciences Center Animal Care and Use Committee.

Insulin autoantibody (IAA) assay

IAs were measured with a 96-well filtration plate micro-IAA assay as described previously²⁷ and expressed as an index. A value of 0.01 or greater is considered positive.

Histology

The pancreata and salivary glands obtained from the mice were fixed in 10% formalin, then embedded in paraffin. Paraffin-embedded tissue sections were stained with haematoxylin and eosin, and parallel sections were also stained with polyclonal guinea-pig anti-insulin or anti-glucagon antibodies (Linco Research Inc.), followed by incubation with a peroxidase-labelled anti-guinea-pig IgG antibody (Kirkegaard & Perry Laboratories). For the scoring of insulinitis, each islet was scored as no insulinitis, peri-islet insulinitis and intra-islet insulinitis by a reader blinded to the categories of the mice. To calculate the insulin-positive area we modified the method of measurement of beta-cell mass in ref. 28. An Olympus BX51 microscope connected to a high-resolution digital camera (Pixera) and the software Image-Pro Plus (Media Cybernetics, Inc.) was used. Pancreas and islet areas were measured in a randomly selected 10–15 views at a final magnification of $\times 400$. Subsequently, islet area and insulin-positive area were measured at $\times 1,000$ magnification. The insulin-positive area divided by the pancreatic area was calculated as (insulin-positive area at $\times 1,000$)/(islet area at $\times 400$)/(pancreatic area at $\times 400$).

Diabetes

Glucose was measured weekly with the FreeStyle blood glucose monitoring system (TheraSense), and the mice were considered diabetic after two consecutive blood glucose values of more than 250 mg dl⁻¹.

Adoptive transfer experiments

Splenocytes (3×10^7) from native insulin-negative NOD mice or standard NOD mice were injected intraperitoneally into 8-week-old NOD.SCID mice and the mice were followed for the development of hyperglycaemia.

Statistics

The peaks of insulin autoantibodies were analysed with Student's *t*-tests. The presence of insulinitis was analysed with the Fisher exact test. Survival curves were analysed with the log-rank test. Statistical tests used PRISM software (Graphpad).

Received 9 January; accepted 8 March 2005; doi:10.1038/nature03523.

- Wegmann, D. R., Norbury-Glaser, M. & Daniel, D. Insulin-specific T cells are a predominant component of islet infiltrates in pre-diabetic NOD mice. *Eur. J. Immunol.* **24**, 1853–1857 (1994).
- Lieberman, S. M. *et al.* Identification of the β cell antigen targeted by a prevalent population of pathogenic CD8⁺ T cells in autoimmune diabetes. *Proc. Natl Acad. Sci. USA* **100**, 8384–8388 (2003).
- Abiru, N. *et al.* Dual overlapping peptides recognized by insulin peptide B:9–23 T cell receptor AV13S3 T cell clones of the NOD mouse. *J. Autoimmun.* **14**, 231–237 (2000).
- Thébault-Baumont, K. *et al.* Acceleration of type 1 diabetes mellitus in proinsulin 2-deficient NOD mice. *J. Clin. Invest.* **111**, 851–857 (2003).
- Moriyama, H. *et al.* Evidence for a primary islet autoantigen (preproinsulin 1) for insulinitis and diabetes in the nonobese diabetic mouse. *Proc. Natl Acad. Sci. USA* **100**, 10376–10381 (2003).
- Jaekel, E., Lipes, M. A. & von Boehmer, H. Recessive tolerance to preproinsulin 2 reduces but does not abolish type 1 diabetes. *Nature Immunol.* **5**, 1028–1035 (2004).
- Daniel, D., Gill, R. G., Sclott, N. & Wegmann, D. Epitope specificity, cytokine production profile and diabetogenic activity of insulin-specific T cell clones isolated from NOD mice. *Eur. J. Immunol.* **25**, 1056–1062 (1995).
- Wong, F. S., Moustakas, A. K., Wen, L., Papadopoulos, G. K. & Janeway, C. A. Jr. Analysis of structure and function relationships of an autoantigenic peptide of insulin bound to H-2K^b that stimulates CD8 T cells in insulin-dependent diabetes mellitus. *Proc. Natl Acad. Sci. USA* **99**, 5551–5556 (2002).
- Simone, E. *et al.* T cell receptor restriction of diabetogenic autoimmune NOD T cells. *Proc. Natl Acad. Sci. USA* **94**, 2518–2521 (1997).
- Mordes, J. P., Bortell, R., Blankenhorn, E. P., Rossini, A. A. & Greiner, D. L. Rat models of type 1 diabetes: genetics, environment, and autoimmunity. *ILAR J.* **45**, 278–291 (2004).
- Todd, J. A., Bell, J. I. & McDevitt, H. O. HLA-DQ β gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **329**, 599–604 (1987).
- Ellerman, K. E. & Like, A. A. Susceptibility to diabetes is widely distributed in normal class IIu haplotype rats. *Diabetologia* **43**, 890–898 (2000).
- Kubosaki, A. *et al.* Targeted disruption of the IA-2 β gene causes glucose intolerance and impairs insulin secretion but does not prevent the development of diabetes in NOD mice. *Diabetes* **53**, 1684–1691 (2004).
- Jaekel, E., Klein, L., Martin-Orozco, N. & von Boehmer, H. Normal incidence of diabetes in NOD mice tolerant to glutamic acid decarboxylase. *J. Exp. Med.* **197**, 1635–1644 (2003).
- Kash, S. F., Condie, B. G. & Baekkeskov, S. Glutamate decarboxylase and GABA in pancreatic islets: lessons from knock-out mice. *Horm. Metab. Res.* **31**, 340–344 (1999).
- Birk, O. S. *et al.* A role of Hsp60 in autoimmune diabetes: analysis in a transgenic model. *Proc. Natl Acad. Sci. USA* **93**, 1032–1037 (1996).
- Kubosaki, A., Miura, J. & Notkins, A. L. IA-2 is not required for the development of diabetes in NOD mice. *Diabetologia* **47**, 149–150 (2004).
- DiLorenzo, T. P. *et al.* During the early prediabetic period in NOD mice, the pathogenic CD8⁺ T-cell population comprises multiple antigenic specificities. *Clin. Immunol.* **105**, 332–341 (2002).
- Mathis, D. & Benoist, C. Back to central tolerance. *Immunity* **20**, 509–516 (2004).
- Peterson, J. D. & Haskins, K. Transfer of diabetes in the NOD-scid mouse by CD4 T-cell clones. Differential requirement for CD8 T-cells. *Diabetes* **45**, 328–336 (1996).
- French, M. B. *et al.* Transgenic expression of mouse proinsulin II prevents diabetes in nonobese diabetic mice. *Diabetes* **46**, 34–39 (1996).
- Mukherjee, R., Chaturvedi, P., Qin, H. Y. & Singh, B. CD4⁺CD25⁺ regulatory T cells generated in response to insulin B:9–23 peptide prevent adoptive transfer of diabetes by diabetogenic T cells. *J. Autoimmun.* **21**, 221–237 (2003).
- Chatenoud, L., Primo, J. & Bach, J. F. CD3 antibody-induced dominant self tolerance in overtly diabetic NOD mice. *J. Immunol.* **158**, 2947–2954 (1997).
- Bluestone, J. A. & Tang, Q. Therapeutic vaccination using CD4⁺CD25⁺ antigen-specific regulatory T cells. *Proc. Natl Acad. Sci. USA* **101** (suppl. 2), 14622–14626 (2004).
- Duvillier, B. *et al.* Phenotypic alterations in insulin-deficient mutant mice. *Proc. Natl Acad. Sci. USA* **94**, 5137–5140 (1997).
- Nakayama, M. *et al.* Establishment of native insulin-negative NOD mice and the methodology to distinguish specific insulin knockout genotypes and a B:16 alanine preproinsulin transgene. *Ann. NY Acad. Sci.* **1037**, 193–198 (2004).
- Yu, L. *et al.* Early expression of antiinsulin autoantibodies of humans and the NOD mouse: evidence for early determination of subsequent diabetes. *Proc. Natl Acad. Sci. USA* **97**, 1701–1706 (2000).
- Sreenan, S. *et al.* Increased β -cell proliferation and reduced mass before diabetes onset in the nonobese diabetic mouse. *Diabetes* **48**, 989–996 (1999).

Acknowledgements This work was supported by grants from the NIH, the Diabetes Endocrine Research Center, the Autoimmunity Prevention Center, the Immune Tolerance Network, the American Diabetes Association, the Juvenile Diabetes Foundation and the Children's Diabetes Foundation. M.N. was supported by a fellowship from the Juvenile Diabetes Foundation and the Naito Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.S.E. (george.eisenbarth@uchsc.edu).

Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope

Sally C. Kent¹*, Yahua Chen¹*, Lisa Bregoli¹, Sue M. Clemmings², Norma Sue Kenyon³, Camillo Ricordi³, Bernhard J. Hering² & David A. Hafler¹

¹Center for Neurologic Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

²Diabetes Institute for Immunology and Transplantation, Department of Surgery, University of Minnesota, Minneapolis, Minnesota 55455, USA

³Diabetes Research Institute, Cell Transplant Center, University of Miami, Miami, Florida 33136, USA

* These authors contributed equally to this work

In autoimmune type 1 diabetes, pathogenic T lymphocytes are associated with the specific destruction of insulin-producing β -islet cells^{1,2}. Identification of the autoantigens involved in triggering this process is a central question. Here we examined T cells from pancreatic draining lymph nodes, the site of islet-cell-specific self-antigen presentation³. We cloned single T cells in a non-biased manner from pancreatic draining lymph nodes of subjects with type 1 diabetes and from non-diabetic controls. A high degree of T-cell clonal expansion was observed in pancreatic lymph nodes from long-term diabetic patients but not from control subjects. The oligoclonally expanded T cells from diabetic subjects with DR4, a susceptibility allele for type 1 diabetes⁴, recognized the insulin A 1–15 epitope restricted by DR4. These results identify insulin-reactive, clonally expanded T cells from the site of autoinflammatory drainage in long-term type 1 diabetics, indicating that insulin may indeed be the target antigen causing autoimmune diabetes.

One of the major issues in understanding the pathophysiology of human autoimmune diseases is to identify the target antigens that drive the clonal expansion of autoreactive T cells. One difficulty in identifying potentially pathogenic T cells and their target antigens arises from observations in experimental animal models of autoimmune disease, which indicate that these pathogenic autoreactive T cells become activated and undergo clonal expansion in the draining lymph nodes of the inflamed target organ^{5–8}. In human autoimmune disease, T cells from the inflammatory sites in rheumatoid arthritis and Graves' disease have been shown to recognize autoantigens^{9,10}. Although autoreactive T cells may enter the circulation for short periods of time, it has not been possible to identify T cells undergoing local clonal expansion in peripheral blood¹¹.

Type 1 diabetes is an autoimmune disease in which auto-destructive T cells infiltrate the pancreatic islets and specifically destroy insulin-producing β -islet cells^{1,2,12}. Studies using the non-obese diabetes (NOD) model of autoimmune diabetes have shown

that dendritic cells migrate from the pancreas to the pancreatic draining lymph nodes, where islet-cell antigens are presented to the pathogenic T cells that mediate the disease³. These investigations led to a study in which clonally expanded T cells were cloned directly from the draining lymph nodes of NOD mice and examined for antigen reactivity. Notably, these investigations revealed that insulin (an islet-cell-specific antigen) was the major target antigen for CD8⁺ T cells in this mouse model of diabetes¹³.

Together, these experiments indicated that clonally expanded, autoreactive T cells could be cloned directly from the pancreatic draining lymph nodes of NOD mice. However, the adaptation of this experimental approach to humans with autoimmune diabetes presents almost insurmountable problems. First, fresh draining lymph nodes from patients with autoimmune diabetes would need to be obtained, and single-cell lymphocyte cloning would have to be performed with relatively high efficiency. Second, unlike in the NOD model, humans will almost always have received insulin injections, which could potentially affect the results. Finally, owing to the long time course of the disease and the rarity of obtaining tissue, it will be virtually impossible to perform any type of kinetic analysis of the T-cell response to self-antigens from draining lymph nodes.

With these caveats in mind, pancreatic draining lymph nodes were obtained from three subjects with autoimmune diabetes and three non-autoimmune subjects. The disease state, age, gender and human leukocyte antigen (HLA) information for each subject are summarized in Table 1. Among the three control subjects, two are normal controls with no detectable immunologic defects, and one is a subject with long-standing type 2 diabetes. Two out of three pancreatic lymph node (PLN) samples (diabetic subjects 1 and 2) were from subjects with long-term disease (Table 1). Both of these subjects expressed the DRB1*0401 and DRB1*0301 class II major histocompatibility complex (MHC) alleles, and one expressed the DQ*0302 class II MHC allele, which are the three alleles highly associated with type 1 diabetes^{4,14}. Subject 3 had a recent onset of clinical disease (1.5 yr). Immunohistochemical examination of the pancreatic tissue sample from this subject showed the remnants of islets, with lymphocytic infiltrate and CD4⁺ T cells within the islets (see Supplementary Fig. 1). This is a rare case compared with most diabetic pancreases, in which usually no islets or T-cell infiltrates are left at the time of organ examination many years after disease onset. For example, examination of the pancreas from long-term diabetic subject 2 showed no infiltrate or cellular islet structure (Supplementary Fig. 2).

A total of 515 independent T-cell clones were generated by single-cell cloning with mitogen in the presence of a blocking anti-Fas antibody. The single-cell cloning efficiency varied between 5% and 20% of the single T cells plated. Polymerase chain reaction with reverse transcription (RT-PCR) was performed for each of the T-cell clones to determine the degree of clonal expansion, and to ask specifically whether the T cells cloned from the draining lymph nodes showed characteristics of T cells undergoing antigen-driven clonal expansion. All of the T cell clones isolated from the three

Table 1 Subject information

Subjects	Age (yr)	Gender	Disease duration (yr)	HLA type
Control				
1	64	M	n.a.	A2, A23 B62, B49 DRB1*0101, *1101 DQB1*0301, *05
2	52	F	n.a.	A1, A3 B7, B37 DR1, DR2
3	45	F	n.a.	A2, A24 B7, B62 DRB1*0401, DR9
Diabetic				
1	39	F	29	A2, A2 B49, B50 DRB1*0401, *0301 DQ 2, DQ3
2	24	F	15	A1, A30, B18, B57 DRB1*0401, *0301 DQB1*0302, *0201
3	27	M	1.5	A3, A24 B7, B35 DRB1*0101, *0101 DQB1*05, *05

Control and diabetic pancreatic draining lymph nodes were from multi-organ donors except for diabetic sample 1, which was collected from a type 1 diabetic undergoing pancreas transplant removal. All samples were obtained with approval for research. Control sample 1 is from a type 2 diabetic. All lymph nodes were processed into single-cell suspensions and frozen in 10% dimethylsulphoxide in FBS within 72 h of harvest. DQ tissue typing was not performed for all samples. n.a., not applicable.

Table 2 T-cell clones isolated from control subjects express few identical V β chains

Control subject	No. of T-cell clones sequenced	No. of identical sequenced V β chains
1	19	2; 2; 2
2	26	2; 2
3	22	2

T-cell clones were isolated from control subject pancreatic draining lymph nodes, and T-cell receptor V β chains were sequenced. In control subject 1, three sets of two identical V β chains were found from 19 clones. In control subject 2, two sets of two identical V β chains were found in 26 clones. In control subject 3, two identical V β chains were seen in 22 T-cell clones. In all three subjects, each V β chain was represented in more than one clone in approximately 10% of isolated clones.

control subjects express heterogeneous T-cell antigen receptors, with only a small portion (~10%) isolated from a single sample expressing identical V β chains (Table 2). These results indicate that there is little restriction of V β families used by T cells isolated from pancreatic lymph nodes of control subjects, and that these cells are representative of a polyclonal repertoire.

In contrast, T-cell clonal expansion was clearly observed in pancreatic lymph nodes of diabetic subjects 1 and 2 (Table 3 and Supplementary Tables 1 and 2), where over half of the T-cell clones expressed identical V β chains. T-cell clones isolated from diabetic subject 3 showed heterogeneous T-cell receptor (TCR) β chains. Analysis of TCR α chains of T-cell clones from diabetic subjects 1 and 2 expressing identical TCR β chains revealed that almost half of the T-cell clones also expressed identical TCR α chains, suggesting that these clones were expanded from the same progenitor cell. Another three clones from diabetic subject 2 also expressed identical TCR α chains. The remaining sequenced clones had different TCR α chains.

To confirm whether T cells with common TCR sequences were derived from the same progenitor T cells in the thymus, rearrangement of TCR γ chains was examined in a panel of clones with identical TCR β sequences. Oligoclonally expanded T cells from diabetic subject 2 containing identical TCR V β and V α sequences also possessed identical TCR V γ chain sequences, demonstrating that these cells are of the same lineage (data not shown).

A series of β -islet-associated candidate T-cell antigens have been identified, including β -cell-specific insulin in humans and NOD mice^{15–18}. These candidates are potentially important in the autoimmune process and are of putative predictive clinical value with

respect to type 1 diabetes. We used these observations to examine the antigen specificity of the clonally expanded T cells: MHC-matched, transformed B cell lines were cultured with peptides from candidate antigens and a panel of T-cell clones derived from the draining lymph nodes.

Clonally expanded CD4⁺ clones from the two long-term diabetic subjects, but not from a control subject, recognized the insulin A 1–15 peptide (Fig. 1a–f). This recognition was specific, as the T-cell clones did not recognize insulin A 5–21 or any other epitope of proinsulin, or glutamic acid decarboxylase 65 (GAD65) peptide 274–286, GAD65 peptide 555–567 or a myelin basic protein (MBP) peptide (residues 85–99) that binds DRB1*1501 and DRB1*0401 (refs 19–21). The amount of insulin peptide required to induce proliferation and cytokine secretion followed a standard dose-response curve. T-cell clones responded to insulin A 1–15 and anti-CD3 stimulation with thymidine incorporation and interleukin (IL)-13 secretion, but no interferon (IFN)- γ secretion (not shown). It should be noted that cytokine secretion from long-term, mitogen-stimulated human T-cell clones in culture might not reflect *in vivo* secretion. Therefore, it is not possible to extrapolate the *in vivo* functional programme of these insulin-reactive T cells from these data.

Two of the subjects with diabetes and one control subject inherited the DRB1*0401 genotype. The other two control subjects and the subject with diabetes for 1.5 yr inherited the DRB1*0101 genotype (DR1). We were unable to expand any T-cell clones from diabetic subject 3, and analysis of antigen reactivity awaits construction of TCR-expressing hybridomas²¹. The T-cell clones investigated from diabetic subjects 1 and 2 expressed CD4. In these experiments, monoclonal antibodies that block MHC DR, but not DQ presentation of antigen, inhibited thymidine incorporation and cytokine secretion in response to insulin peptide (Fig. 1b–f). Use of cell lines homozygous for HLA*DRB1 alleles showed that the insulin-reactive T cells were DRB1*0401- and not DRB1*0301-restricted in the diabetic subjects expressing those genotypes (Fig. 2a). This is of interest because DR4 is strongly associated with the incidence of type 1 diabetes⁴.

To be certain that the insulin A 1–15 peptide itself was non-mitogenic and that peptide reactivity was specific, more T-cell clones and peptides were tested. T-cell clones from diabetic subjects 1 and 2 responded only to insulin A 1–15 (in a dose-dependent

Table 3 Clonally expanded T cells isolated from subjects with type 1 diabetes

V β J β usage	V β /NDN/J β junctional sequence*	Clone frequency	V α J α usage	V α /N/J α junctional sequence	Clone frequency
Diabetic 1.Mi V β 29-1*03 J β 2-3*01 V β 7-9*03 J β 2-3*01	DSSIYLCSS / VEATRA / DTQYFGPGTRLTVL DSAMYLCASS / LAVIR / TDTQYFGPGTRLTVL	(10/20) 50% (2/20) 10%	V α 8-3*02 J α 44*01 n.d.	YFCA / VGALA / GTASKLFTGTGTRL	(5/10) 25%
Diabetic 2.Ba V β 5-1*01 J β 2-3*01	DSALYLCASSL / ATSGGGS / DTQYFGPGTRL	(14/27) 51.8%	V α 39*01 J α 33*01 V α 22*01 J α 52*01 V α 26-2*01 J α 47*01	YFCA / VWNM / DSNYQLIWGAGTKL YFCA / DAGGTSYKL / FGQGTL YYCI / PGSEE / YGNKLVFGAGTL	(7/14) 25.9% (3/14) 11.0% (1/14) 3.7%
Diabetic 3 V β 4-1*02 J β 2-1*01 V β 5-1*01 J β 1-1*01 V β 14*01 J β 1-1*01 V β 28*01 J β 2-3*01	DSALYLCASSQ / VRLAGGG / EQFFGPGTRLTV DSALYLCASSL / GQGE / TEAFFGQGTRL DSGVYFCAS / RNLGL / NTEAFFGQGARL HTSMYLCASS / FRRV / TDTQYFGPGTRL	(2/43) 4.65% (2/43) 4.65% (2/43) 4.65% (2/43) 4.65%			

*NDN: N, N-region additions; D, diversity region additions.

From diabetic subject 1, 50% of T-cell clones expressed identical V β chains, and of these clones, five expressed identical V α rearrangements (25%). From diabetic subject 2, 14/27 (51.8%) of T-cell clones expressed identical rearranged V β chains, and of these 14, seven were identical for TCR V α (25.9%). Another four clones used the same V β chain, but were paired with two other V α chains. From diabetic subject 3, V β chains were expressed at low frequency (4.65%), and each of these sets of clones expressed different V α chains (not shown). n.d., not determined.

manner) and not to other peptides reported to bind to DRB1*0401 (Fig. 2b, c). Additionally, T-cell clones derived from the cerebrospinal fluid of a patient with multiple sclerosis (DRB1*0401, DRB1*0301) or clones from an DRB1*0401 subject named NC3 (including one of the expanded clones, NC3.13; see Table 2) did not recognize the peptide when presented by the MHC DRB1*0401 allele (Fig. 2d). T-cell clones from diabetic subject 2 with the same V_{β} chain but a different β -chain CDR3 region and different V_{α} chain from the clones reported in Fig. 1 did not respond to the insulin A 1–15 peptide presented by the MHC DRB1*0401 allele (Clone Ba.32, Fig. 2d). A T-cell clone recognizing the self-antigen MBP was examined, and this clone did not respond to the insulin peptide by proliferation or cytokine secretion (Fig. 2e).

NOD mouse studies have generally examined islet-infiltrating lymphocyte, draining lymph node or splenic cell populations for

TCR repertoires in mice with recent disease onset or established disease^{22–25}, but not at the equivalent long-term point of diabetic subjects 1 and 2 because NOD mice are not generally maintained on exogenous insulin for extended periods after disease onset. In contrast, diabetic subject 3 had a relatively shorter disease duration of 1.5 yr. For this subject, there were scattered $CD4^{+}$ T lymphocyte infiltrates in the pancreatic islets (Supplementary Fig. 1). As one might suspect, the timing of a TCR repertoire assay after disease onset is crucial. Examination of the TCR repertoire in islet infiltrates of 2–3-week-old NOD mice showed enrichment of a few TCR α and β sequences with similar sequences to those seen in the pancreatic draining lymph nodes²⁶. However, there are a number of reports indicating a heterogeneous TCR repertoire in islet-infiltrating cells, spleen and lymph nodes of 7–11-week-old NOD mice^{23–25}.

In the diabetic subjects, we found MHC-restricted recognition of the insulin A 1–15 peptide with $CD4^{+}$ clones that represented almost half of the clonable T cells in the draining pancreatic lymph nodes. This allowed us to calculate a minimal frequency of insulin A 1–15 antigen-reactive T cells of 1.04% for diabetic subject 1 and 0.86% for diabetic subject 2, on the basis of the cloning frequency for each lymph node. Although we do not know the autoantigen(s) potentially driving any clones not reactive to insulin and the cells we could not clone, this nevertheless represents an extremely high frequency of antigen-reactive $CD4^{+}$ cells generated without *in vitro* exposure to antigen.

There are a number of caveats to be considered in the interpretation of our presented clinical data. First, it is important to determine whether the clonally expanded, insulin-A-1–15-reactive $CD4^{+}$ cells in the pancreatic draining lymph nodes were found specifically in the draining lymph nodes. In addition, the diabetic subjects had high blood glucose levels and were subjected to insulin injections over a long period of time; it is possible that daily administration of exogenous insulin could initiate and sustain a systemic anti-insulin T-cell response.

A series of experiments was performed to examine these questions. As the clonally expanded $CD4^{+}$ cells from diabetic subject 2 expressed the $V_{\beta}5.1$ TCR, we either plated $CD4^{+}V_{\beta}5.1^{+}$ T cells from the subject's spleen at a density of one cell per well and expanded them in the presence of irradiated allogeneic peripheral blood mononuclear cells (PBMCs), mitogen and IL-2, or instead directly amplified the $V_{\beta}5.1$ CDR3 sequence from the single-cell $CD4^{+}V_{\beta}5.1^{+}$ genomic DNA and sequenced the PCR product (see Supplementary Table 3). None of the four independent $CD4^{+}V_{\beta}5.1^{+}$ clones recognized any pro-insulin peptide (data not shown). Moreover, out of the 70 independent TCR V_{β} sequences obtained from 160 wells, none showed similarity to the expanded, insulin-reactive T-cell clone Ba.14, found in the pancreatic draining lymph nodes of this long-term type 1 diabetic subject. We also examined the panel of T-cell clones generated from the pancreatic draining lymph nodes from NC1, a long-term type 2 diabetic subject with dysregulated, non-autoimmune insulin responses. Using autologous lymph node cells as antigen-presenting cells, none of the T-cell clones from this subject's pancreatic lymph node recognized any peptide of proinsulin (Fig. 2f).

The lack of a systemic response to insulin peptides, together with the absence of the potentially pathogenic T-cell clones in the spleen of the diabetic subject, suggest that the parenteral administration of insulin is not the explanation for the pancreatic lymph-node-specific responses. Moreover, it has been demonstrated that for islet-cell transplants in patients with no insulin secretion (similar to the diabetic subjects in this investigation), there is a highly pathogenic T-cell response to islet cells. Their transplantation, even into an identical twin, results in tissue destruction with lymphocytic infiltrate within days of engraftment²⁷. The clinical transplant data provide strong evidence that there is persistence of T cells, probably residing in the pancreatic draining lymph nodes, and (considering the kinetics of the islet cell rejection) capable of recognizing islet-cell

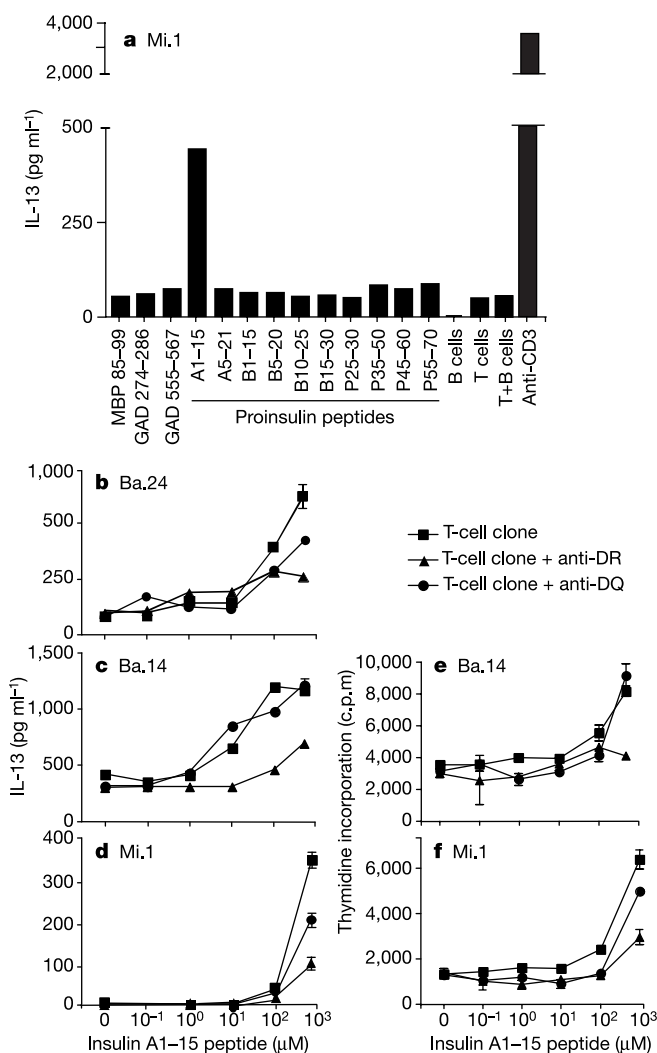


Figure 1 Expanded T-cell clones from diabetic PLN recognize insulin A 1–15 presented by the MHC DRB1*0401 allele. **a**, T-cell clone Mi.1 from diabetic subject 1 responded to insulin A 1–15 peptide in the context of Priess B cells (homozygous for DRB1*0401) but not other proinsulin peptides or other peptides reported to bind to DRB1*0401. All peptides were present at 250 μM. **b–f**, Clones Ba.24 (**b**), Ba.14 (**c**, **e**) and Mi.1 (**d**, **f**) recognize insulin A 1–15 peptide in a dose-dependent manner, as shown by IL-13 secretion levels (**b–d**) and by thymidine incorporation (**e**, **f**). This insulin-reactivity was blocked by an anti-DR antibody, but not by an anti-DQ antibody. Ba.24 and Ba.14 are expanded clones from diabetic subject 2 and possess identical TCR V_{β} and V_{α} chains. Error bars represent the s.e.m. of triplicates; one representative experiment out of three similar experiments is shown.

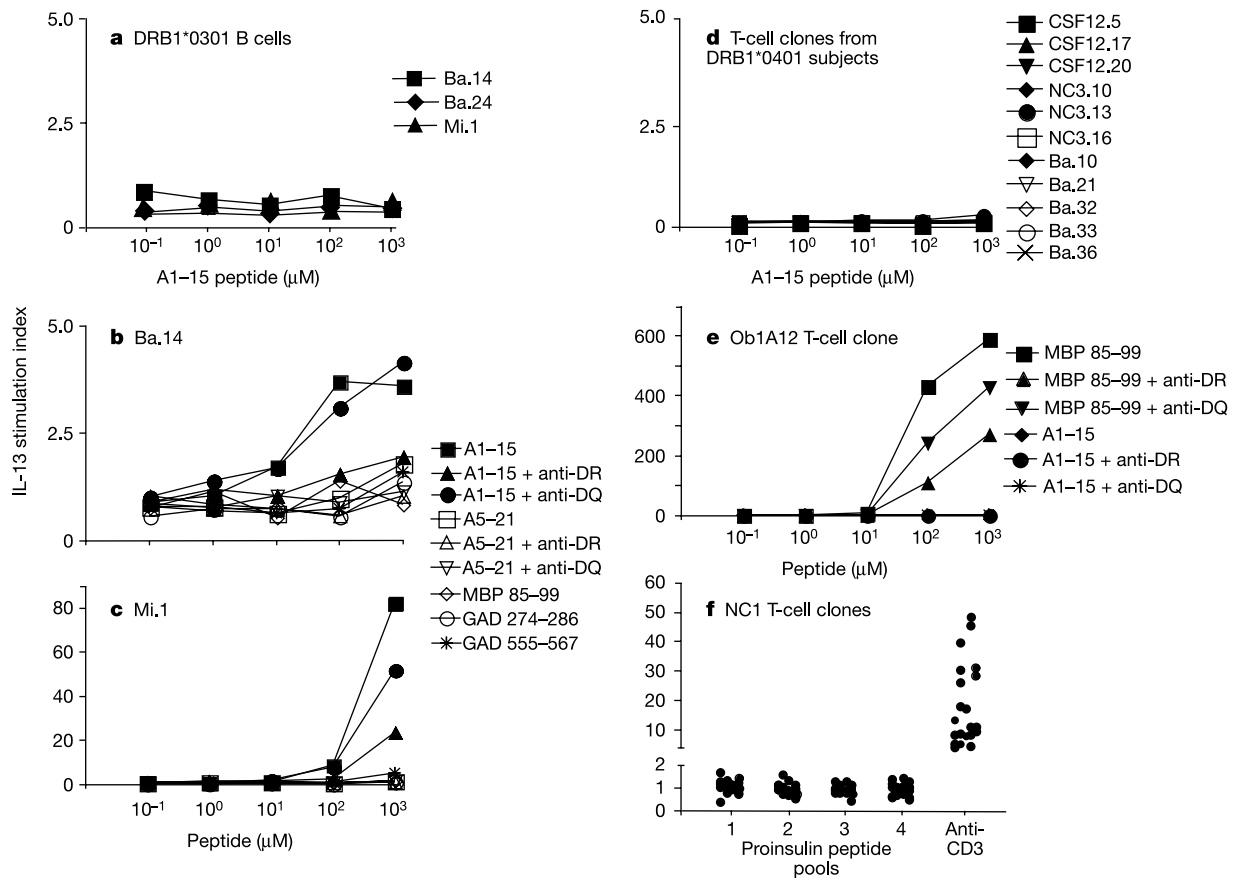


Figure 2 Specificity of peptide- and T-cell-clone-reactivity. **a**, Expanded T-cell clones from diabetic subjects 1 and 2 do not recognize insulin A 1–15 peptide in the context of QBL B cells homozygous for DRB1*0301. **b**, **c**, Expanded T-cell clones Ba.14 (**b**) and Mi.1 (**c**) did not respond to increasing doses of insulin A 5–21, MBP 85–99, or GAD65 peptides 275–286 or 555–567 presented by the MHC DRB1*0401 allele (Priess B cells). **d**, T-cell clones derived from the cerebrospinal fluid of a multiple sclerosis patient (CSF12.5, 12.17 and 12.20) or from the PLN of a DRB1*0401-expressing normal control subject (NC3.10, 3.13, 3.16) did not recognize insulin A 1–15 peptide presented by DRB1*0401. Five non-oligoclonally expanded T-cell clones from diabetic subject 2 were not reactive with insulin A 1–15 peptide, including clone Ba.32 (which has the same V_{β} chain as Ba.14 and Ba.24, but with a different CDR3 region and another V_{α} chain,

$V_{\alpha}25.1$). **e**, A T-cell clone of known specificity derived from the peripheral blood of a multiple sclerosis patient (Ob1A12 recognizing MBP 85–99 presented by DRB1*1501 and DRB1*0401) did not respond to insulin A 1–15 peptide presented by DRB1*0401. **f**, None out of 21 T cell clones (including three expanded clones) from the PLN of the type 2 diabetic subject NC1 (Table 2) recognized pools of proinsulin peptides in the context of autologous pancreatic lymph node cells. However, they did incorporate thymidine and secrete IFN- γ and IL-13 in response to anti-CD3 stimulation (proliferation and IFN- γ secretion not shown). Pool 1, insulin A chain peptides 1–15 and 5–21. Pool 2, insulin B chain peptides 1–15 and 5–20. Pool 3, insulin B chain peptides 10–25 and 15–30. Pool 4, proinsulin-specific peptides P25–30, P35–50, P45–60 and P55–70). Each peptide was present at 250 μ M.

antigens and destroying antigen-expressing cells in patients with long-standing diabetes.

Although we generated over 515 independent T-cell clones, another obvious caveat is that we examined three subjects. Other diabetic subjects might have different antigen reactivities driving the autoimmune response. It should be pointed out that intense investigation of each human subject effectively translates to examination of a single strain of mice. These experiments demonstrate the feasibility of non-biased T-cell cloning from the draining lymph node of an organ targeted in an autoimmune disease in order to identify a putative autoantigen. Our results with humans confirm those found in the NOD mouse model, identifying insulin as a critical autoantigen in diabetes.

Is insulin an important antigen—or perhaps the primary antigen—driving the autoimmune response in patients with type 1 diabetes? In the NOD model of diabetes, recent experiments have shown that gene disruption of insulin prevents the onset of diabetes^{28,29}. In addition, induction of tolerance to pre-proinsulin 2 reduced the onset and degree of diabetes in NOD mice, indicating that insulin and its precursors are important in diabetes initiation. Second, experiments using a similar approach of non-biased T-cell cloning¹³ similarly identified insulin as a critical T-cell antigen in the

NOD model of diabetes. However, other islet antigens, such as GAD65, must be considered as crucial candidates in diabetes initiation³⁰. These data, together with the results presented here, provide perhaps the most conclusive evidence possible from *in vitro* experiments that insulin is a critical antigen in the aetiology of human type 1 diabetes. Ultimately, to fulfill ‘Koch’s postulate’ with regards to identifying autoantigens in T-cell mediated human autoimmune diseases, clinical experiments will be required in which tolerization to the antigen results in disease amelioration. □

Methods

Tissue

Pancreatic draining lymph nodes (PLN, superior and inferior pancreaticoduodenal and supra/infrapancreatic lymph nodes) and spleen were harvested from multi-organ donors or during surgery, with appropriate institutional review.

T-cell cloning and examination of antigen reactivity

Single-cell suspensions were made from PLN and spleen cells within 72 h of organ harvest, and were frozen in 10% dimethylsulphoxide in FBS and stored in liquid nitrogen until use. T cells were cloned from PLN at 0.3 cells per well with 3μ g ml⁻¹ phytohemagglutinin (PHA-P, Remel) and irradiated, allogeneic PBMCs and 20 U ml⁻¹ recombinant human IL-2 (Tecin, NCI) in the presence of 10μ g ml⁻¹ anti-Fas antibody (Boehringer Ingelheim) to prevent death of re-activated T cells from the PLN when activated with allogeneic feeders and PHA. Media for T-cell cultures contained 5% heat-inactivated human male AB

serum (Omega Scientific) in RPMI 1640 with 10 mM HEPES buffer, 2 mM l-glutamine, 100 U ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin (all from Cambrex BioScience). T-cell clones were expanded with IL-2, assayed on day 9 or 10 post-stimulation and re-stimulated as previously described¹¹.

Antigen reactivity was examined using irradiated (5,000 rads) Priess EBV-transformed B cells (homozygous for DRB1*0401) or QBL B cells (homozygous for DRB1*0301) pulsed with peptide (250 µM for Fig. 1a) in the presence or absence of antibody (10 µg ml⁻¹ anti-DR LB3.1 and anti-DQ IVD12) for 2 h, washed and plated in triplicate at 50,000 cells per well with equal numbers of T-cell clones. Each T-cell clone was also plated onto plate-bound anti-CD3 antibody (OKT3; 0.05 µg per well) to assess the viability of each clone in each experiment. After 48 h, 20 U ml⁻¹ IL-2 was added to each well. Supernatants were collected after a further 24 h for measurement by cytokine ELISA (BD PharMingen). For testing of T-cell clones from the type 2 diabetic subject NC1, irradiated, autologous pancreatic draining lymph node cells were pulsed with pools of peptides (each at 250 µM) for 2 h, washed and plated at 120,000 cells per well with 50,000 T cells. The remainder of the assay was as described above.

Sequencing of T-cell receptors

RNAs were isolated using the standard RNeasy method (Tel-Test). Complementary DNAs were synthesized using Superscript Reverse Transcriptase (Gibco) and oligo dT as the primer for reverse transcription. PCR primers for both V_α and V_β families were designed and grouped according to refs 31 and 32. PCR products were purified from agarose gel and subjected to DNA sequencing (Brigham & Women's Hospital Sequencing Facility). The V_α and V_β sequences obtained were aligned against the ImMunoGeneTics (IMGT) database.

Received 14 March; accepted 11 April 2005; doi:10.1038/nature03625.

- Atkinson, M. A. & Maclaren, N. K. The pathogenesis of insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **331**, 1428–1436 (1994).
- Eisenbarth, G. S. Type 1 diabetes mellitus: a chronic autoimmune disease. *N. Engl. J. Med.* **314**, 1360–1368 (1986).
- Hoglund, P. *et al.* Initiation of autoimmune diabetes by developmentally regulated presentation of islet cell antigens in the pancreatic lymph nodes. *J. Exp. Med.* **189**, 331–339 (1999).
- Thomson, G. *et al.* Genetic heterogeneity, modes of inheritance, and risk estimates for a joint study of Caucasians with insulin-dependent diabetes mellitus. *Am. J. Hum. Genet.* **43**, 799–816 (1988).
- Katz, J. D., Wang, B., Haskins, K., Benoist, C. & Mathis, D. Following a diabetogenic T cell from genesis through pathogenesis. *Cell* **74**, 1089–1100 (1993).
- Acha-Orbea, H. & McDevitt, H. O. The role of class II molecules in development of insulin-dependent diabetes mellitus in mice, rats and humans. *Curr. Top. Microbiol. Immunol.* **156**, 103–119 (1990).
- Haskins, K., Portas, M., Bergman, B., Lafferty, K. & Bradley, B. Pancreatic islet-specific T-cell clones from nonobese diabetic mice. *Proc. Natl Acad. Sci. USA* **86**, 8000–8004 (1989).
- Wegmann, D. R., Norbury-Glaser, M. & Daniel, D. Insulin-specific T cells are a predominant component of islet infiltrates in pre-diabetic NOD mice. *Eur. J. Immunol.* **24**, 1853–1857 (1994).
- Londei, M. *et al.* Persistence of collagen type II-specific T-cell clones in the synovial membrane of a patient with rheumatoid arthritis. *Proc. Natl Acad. Sci. USA* **86**, 636–640 (1989).
- Dayan, C. M. *et al.* Autoantigen recognition by thyroid-infiltrating T cells in Graves disease. *Proc. Natl Acad. Sci. USA* **88**, 7415–7419 (1991).
- Hafler, D. A. *et al.* Oligoclonal T-lymphocytes in the cerebrospinal fluid of patients with inflammatory central nervous system diseases. *J. Exp. Med.* **167**, 1625–1644 (1988).
- Roep, B. O. T-cell responses to autoantigens in IDDM. *Diabetes* **45**, 1147–1156 (1996).
- Wong, F. S. *et al.* Identification of an MHC class I-restricted autoantigen in type 1 diabetes by screening an organ-specific cDNA library. *Nature Med.* **5**, 1026–1031 (1999).
- Todd, J. A., Bell, J. I. & McDevitt, H. O. HLA antigens and insulin-dependent diabetes. *Nature* **333**, 710 (1988).
- Schlott, N. C. *et al.* Comparison of cytokine ELISpot assay formats for the detection of islet antigen autoreactive T cells. Report of the third immunology of diabetes society T-cell workshop. *J. Autoimmun.* **21**, 365–376 (2003).
- Peakman, M. *et al.* Characterization of preparations of GAD65, proinsulin, and the islet tyrosine phosphatase IA-2 for use in detection of autoreactive T-cells in type 1 diabetes: report of phase II of the Second International Immunology of Diabetes Society Workshop for Standardization of T-cell assays in type 1 diabetes. *Diabetes* **50**, 1749–1754 (2001).
- Eisenbarth, G. S. *et al.* Insulin autoimmunity: prediction/precipitation/prevention type 1A diabetes. *Autoimmun. Rev.* **1**, 139–145 (2002).
- Harrison, L. C. *et al.* Islet-reactive T cells are a marker of preclinical insulin-dependent diabetes. *J. Clin. Invest.* **89**, 1161–1165 (1992).
- Astill, T. P., Ellis, R. J., Arif, S., Tree, T. I. M. & Peakman, M. Promiscuous binding of proinsulin peptides to Type 1 diabetes-permissive and -protective HLA class II molecules. *Diabetologia* **46**, 496–503 (2003).
- Endl, J. *et al.* Identification of naturally processed T cell epitopes from glutamic acid decarboxylase presented in the context of HLA-DR alleles by T lymphocytes of recent onset IDDM patients. *J. Clin. Invest.* **15**, 2405–2415 (1997).
- Mycko, M. P. *et al.* Cross-reactive TCR responses to self antigens presented by different MHC class II molecules. *J. Immunol.* **173**, 1689–1698 (2004).
- Tian, J., Lehmann, P. & Kaufman, D. Determinant spreading of T helper cell 2 (Th2) responses to pancreatic islet autoantigens. *J. Exp. Med.* **186**, 2039–2043 (1997).
- Nakano, N., Kikutani, H., Nishimoto, H. & Kishimoto, T. T cell receptor V gene usage of islet β cell-reactive T cells is not restricted in non-obese diabetic mice. *J. Exp. Med.* **173**, 1091–1097 (1991).
- Miyazaki, A. *et al.* Predominance of T lymphocytes in pancreatic islets and spleen of pre-diabetic non-obese diabetic (NOD) mice: a longitudinal study. *Clin. Exp. Immunol.* **60**, 622–630 (1985).
- Candeias, S., Katz, J., Benoist, C., Mathis, D. & Haskins, K. Islet-specific T-cell clones from nonobese diabetic mice express heterogeneous T-cell receptors. *Proc. Natl Acad. Sci. USA* **88**, 6167–6170 (1991).
- Baker, F., Lee, M., Chien, Y.-h. & Davis, M. Restricted islet-cell reactive T cell repertoire of early pancreatic islet infiltrates in NOD mice. *Proc. Natl Acad. Sci. USA* **99**, 9374–9379 (2002).
- Sibley, R. K., Sutherland, D. E., Goetz, F. & Michael, A. F. Recurrent diabetes mellitus in the pancreas

iso- and allograft. A light and electron microscopic and immunohistochemical analysis of four cases. *Lab. Invest.* **53**, 132–144 (1985).

- Jaekel, E., Lipes, M. A. & von Boehmer, H. Recessive tolerance to preproinsulin 2 reduces but does not abolish type 1 diabetes. *Nature Immunol.* **5**, 1028–1035 (2004).
- Nakayama, M. *et al.* Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice. *Nature* doi:10.1038/nature03523 (this issue).
- Yoon, J. W. *et al.* Control of autoimmune diabetes in NOD mice by GAD expression or suppression in beta cells. *Science* **284**, 1183–1187 (1999).
- Han, M. *et al.* Invariant or highly conserved TCR α are expressed on double-negative (CD3⁺CD4[−]CD8[−]) and CD8⁺ T cells. *J. Immunol.* **163**, 301–311 (1999).
- Akatsuka, Y., Martin, E., Madonik, A., Barsoukov, A. & Hansen, J. Rapid screening of T-cell receptor (TCR) variable gene usage by multiplex PCR: application for assessment of clonal composition. *Tissue Antigens* **53**, 122–134 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank V. Kuchroo and G. Fathman for critical reading of the manuscript. We thank G. Nepom for B cell lines, and R. Neal Smith and N. Kirchhof for expert immunohistochemical and histological tissue staining. These studies were supported by grants to D.A.H. (NIH), to Y.C. (JDRFI Fellowship), to L.B. (NMSS Fellowship) and to S.C.K. (Boston Area Diabetes Endocrinology Research Center). D.A.H. is a recipient of the NIH Javits Investigator Award.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.A.H. (dhafler@rics.bwh.harvard.edu).

Enhancement of cellular memory by reducing stochastic transitions

Murat Acar, Attila Becskei & Alexander van Oudenaarden

Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

On induction of cell differentiation, distinct cell phenotypes are encoded by complex genetic networks^{1–3}. These networks can prevent the reversion of established phenotypes even in the presence of significant fluctuations. Here we explore the key parameters that determine the stability of cellular memory by using the yeast galactose-signalling network as a model system. This network contains multiple nested feedback loops. Of the two positive feedback loops, only the loop mediated by the cytoplasmic signal transducer Gal3p is able to generate two stable expression states with a persistent memory of previous galactose consumption states. The parallel loop mediated by the galactose transporter Gal2p only increases the expression difference between the two states. A negative feedback through the inhibitor Gal80p reduces the strength of the core positive feedback. Despite this, a constitutive increase in the Gal80p concentration tunes the system from having destabilized memory to having persistent memory. A model reveals that fluctuations are trapped more efficiently at higher Gal80p concentrations. Indeed, the rate at which single cells randomly switch back and forth between expression states was reduced. These observations provide a quantitative understanding of the stability and reversibility of cellular differentiation states.

Complex gene and protein networks store cellular memory by creating two or many discrete, stable states of network activity^{4–6}. The generation of bistability by simple feedback loops in synthetic circuits is well understood^{7–10}. However, naturally occurring networks, in particular in eukaryotic organisms, possess a complex organization of multiple nested feedback loops, making an analysis of system dynamics disproportionately more complicated^{11,12}. These networks are exemplified by the galactose signalling pathway

in the yeast *Saccharomyces cerevisiae*. Despite extensive data on its molecular interactions, a prediction of its dynamical system behaviour a priori is challenging^{13,14}. The galactose signal propagates through a four-stage signalling cascade. At the uppermost stage is Gal2p, which imports extracellular galactose into the cell. Subsequently, intracellular galactose binds to and activates Gal3p (refs 15, 16). At the third stage of this cascade, the activated Gal3p binds to and sequesters Gal80p in the cytoplasm, depleting Gal80p from the nucleus¹⁷. The transcriptional activator Gal4p, which is constitutively bound to promoters of the *GAL* genes¹⁸, is then released from the inhibitory action of Gal80p and activates expression of genes at the output of the cascade, including *GAL1*, *GAL2*, *GAL3* and *GAL80*. Because an increase in Gal2p and Gal3p concentration results in enhanced transcriptional activity, these proteins close two positive feedback loops. The opposite holds for Gal80p, which is part of a negative feedback loop. To read out the Gal4p activity in single yeast cells we monitored the expression of yellow fluorescent protein (YFP) driven by the *GAL1* promoter (Fig. 1).

Because the *GAL* regulatory network contains two positive feedback loops, this network has the potential for exhibiting multistability. A convenient experimental way to probe for multistability is to subject the network to different initial conditions and explore whether the network gets locked in different stable expression states. We therefore grew wild-type cells for 12 h either in the absence of galactose ('raffinose history') or in the presence of 2% galactose ('galactose history') and subsequently for a further 27 h at various concentrations of galactose. Raffinose was chosen as a neutral sugar because it neither activates nor represses the galactose-regulated genes¹⁹. To maintain a constant concentration of galactose in the culture medium, cells were grown at low densities so that galactose depletion was negligible (Supplementary Table S2). In Fig. 2a expression histograms are shown for wild-type cells for the two initial conditions and several galactose concentrations. At low galactose concentrations, cells have a basal expression of YFP reflecting low Gal4p activity, whereas at larger galactose

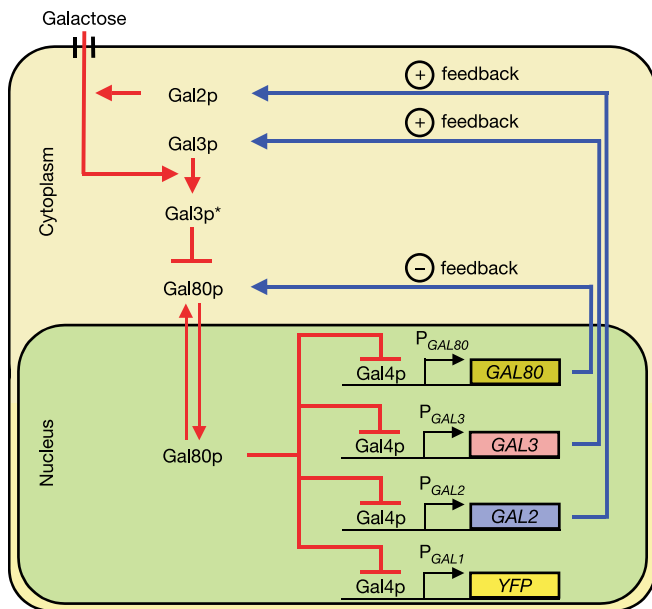


Figure 1 The galactose signalling pathway. Red arrows denote the four-stage signalling cascade in which the external galactose signal controls the transcriptional activity of the *GAL* genes. The galactose-bound state of Gal3p is denoted by Gal3p*. Pointed and blunt arrows reflect activation and inhibition, respectively. The double red arrows represent shuttling of Gal80p between the cytoplasm and the nucleus. The blue arrows denote feedback loops established by Gal2p, Gal3p and Gal80p.

concentrations YFP expression is more than 100-fold higher, reflecting high Gal4p activity. The response to the two different initial conditions depends strongly on the galactose concentration. At low (less than 0.012%) and high (more than 0.35%) galactose concentrations the expression distributions after 27 h do not depend on the history, and typically reach a steady state after 6 h. We classify this behaviour as history independent (absence of memory) because the system approaches the same unique expression distribution when coming from different initial conditions. However, for intermediate galactose concentrations the expression distributions obtained from the two different histories are significantly different and the system displays a memory of the initial galactose consumption state. We classify this behaviour as history dependent (persistent memory) because cells become stably locked into two different expression states for periods much longer than the history-independent system would need to reach steady state. Similar behaviour is obtained when different promoters with Gal4p-binding sites are used (Fig. 3a) to drive YFP expression. In addition, a strong correlation between promoter activities is observed when the activities of two different *GAL* promoters are simultaneously monitored in single cells (Supplementary Fig. S1). This implies that the transcriptional activity of the *GAL1* promoter, and therefore YFP fluorescence, is a faithful reporter for Gal4p activity.

To pinpoint the network interaction responsible for the persistent memory we systematically interrupted the three feedback loops. We found that the positive feedback loop mediated by Gal2p was not

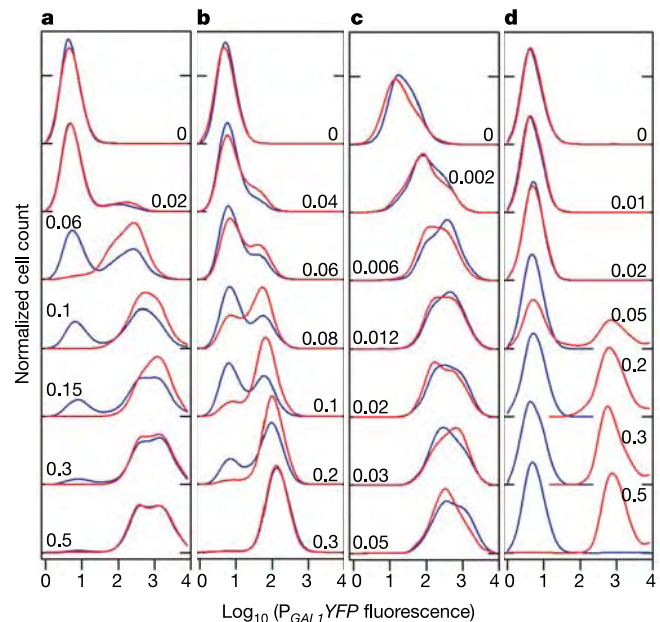


Figure 2 History-dependent experiments reveal the regulatory interaction necessary for persistent memory. Expression distributions were obtained from at least 10,000 cells and are plotted as a function of $\log_{10}(\text{YFP fluorescence})$. Day-to-day variation in the expression histograms is typically less than 5%. Blue and red distributions denote cells that were initially grown for 12 h without galactose (but with 2% raffinose) and with 2% galactose (and 2% raffinose), respectively. After this initial incubation cells were grown for a further 27 h in various concentrations of galactose as specified. **a**, Wild-type strain MA0207. **b**, $\Delta gal2$ strain MA0215. The intensity of the high state is about 50-fold smaller than wild type. **c**, *GAL3* loop knockout (MA0182). A doxycycline concentration of $0.05 \mu\text{g ml}^{-1}$ was used. For this concentration the Gal3p expression corresponds to 80% of the Gal3p level observed in a wild-type strain induced with 0.5% galactose (Supplementary Fig. S2). **d**, *GAL80* loop knockout (MA0188). A doxycycline concentration of $0.05 \mu\text{g ml}^{-1}$ was used. For this concentration the Gal80p expression is very similar to the Gal80p level observed in a wild-type strain induced with 0.5% galactose (Supplementary Fig. S4).

necessary for memory storage (Fig. 2b). *gal2Δ* cells still display history-dependent behaviour, although the expression difference between the two states is reduced. In this case galactose is taken up by non-essential sugar transporters²⁰. The non-essential nature of this feedback loop for memory storage contrasts with prokaryotic metabolic networks, in which positive feedback through sugar transporters often defines the dynamics of the system²¹.

To explore the role of the feedback loops through *GAL3* and *GAL80* we constructed loop knockouts in which the feedback loop was interrupted by replacement of the endogenous, Gal4p-dependent, promoter by an externally inducible, Gal4p-independent, P_{TET} promoter. The capacity of storing the initial galactose consumption state was abolished by disrupting the *GAL3* loop. Irrespective of initial conditions, the cell population approached the same unique distribution (Fig. 2c). The mean expression level increased gradually with increasing galactose concentration, in contrast to the discrete transition observed for the wild-type and *gal2Δ* cells (Fig. 2a, b). This behaviour was observed for the entire *GAL3* expression range examined (from 5% to 300% with respect to wild-type Gal3p levels; Supplementary Fig. S2) and shows that feedback through Gal3p is necessary for memorizing the initial metabolic state in the wild-type network. A slow response of *GAL* genes has been observed in strains lacking a functional *GAL3* gene, several days after induction with galactose^{22,23}. The effect of this long-term adaptation depends strongly on which medium or carbon source is used. In our medium, with raffinose always present as a carbon source, long-term adaptation is not observed within 120 h after the addition of galactose to the medium and is therefore not relevant for the memory experiments (Supplementary Fig. S3).

When Gal80p was expressed constitutively at concentrations comparable to those in the wild type, the range of galactose concentration over which this system displayed persistent memory was significantly widened in comparison with wild-type cells (Fig. 2d). This is consistent with the earlier observation that the positive feedback through Gal3p is necessary for memory storage. In wild-type cells the negative feedback through Gal80p effectively weakens the effect of the positive feedback through Gal3p and therefore decreases the persistent memory region in comparison with the *GAL80* loop knockout in which the negative feedback is abolished. For a high concentration of galactose (more than 0.1%) the expression distributions after 27 h are indistinguishable from the initial conditions, showing extreme persistence of the initial expression states. For example, cells initially grown in the absence of galactose still exhibited the same low Gal4p activity level even after 27 h of growth in 0.5% galactose. At this galactose concentration wild-type cells would display maximum Gal4p activity (Fig. 2a). In this regime the initial condition determines the future expression state, not the current concentration of galactose in the medium.

This system was systematically explored over a broad range of *GAL80* expression (Fig. 3b, Supplementary Fig. S4). In a large part of parameter space (galactose versus *GAL80* expression) the system displays persistent memory bordered by regions in which system memory is absent. In the latter case the system either approaches a state of low Gal4p activity (OFF) or high Gal4p activity (ON) independently of the history. Interestingly, at lower *GAL80* expression level a small region in parameter space was identified in which system behaviour displayed destabilized memory. In this region the expression distribution displayed two distinct peaks; however, the system response was not history dependent. We proposed that although the system has two stable states, fluctuations in the gene expression of key regulators drive random transitions between the two states, resulting in a destabilization of the memory. Indeed, a significant cell-to-cell variation of gene expression levels was observed for several promoters in budding yeast^{24,25}.

To reveal the effect of these fluctuations on the memory, the stability of the expression states was quantified by using the concept

of energy landscapes^{26–28}. The first-order differential equation describing the time evolution of the Gal3p concentration is analogous to the equation of motion of an overdamped particle in an energy landscape, where the particle position is analogous to the concentration of Gal3p. This analogy provides an intuitive representation for the stability of cellular expression states because minima in the energy landscape correspond to stable states that are separated by an energy barrier (Fig. 3c). Fluctuations in gene expression are naturally introduced by analogy with the temperature experienced by the particle in the energy landscape. A particle trapped in an energy well at zero temperature will never escape; however, at elevated temperatures the particle will undergo thermally activated transitions across the barrier. In this case the escape rate is proportional to $\exp(-\Delta U/k_B T)$, where ΔU represents the energy barrier, k_B the Boltzmann constant and T the absolute temperature²⁹. The larger the energy barrier, the more efficiently the fluctuations are trapped in the vicinity of the stable states. The *GAL* system is characterized by two energy barriers: the barrier $\Delta U_{\text{OFF} \rightarrow \text{ON}}$ that an OFF cell has to overcome to switch ON, and the barrier $\Delta U_{\text{ON} \rightarrow \text{OFF}}$ for the opposite transition. The barrier height

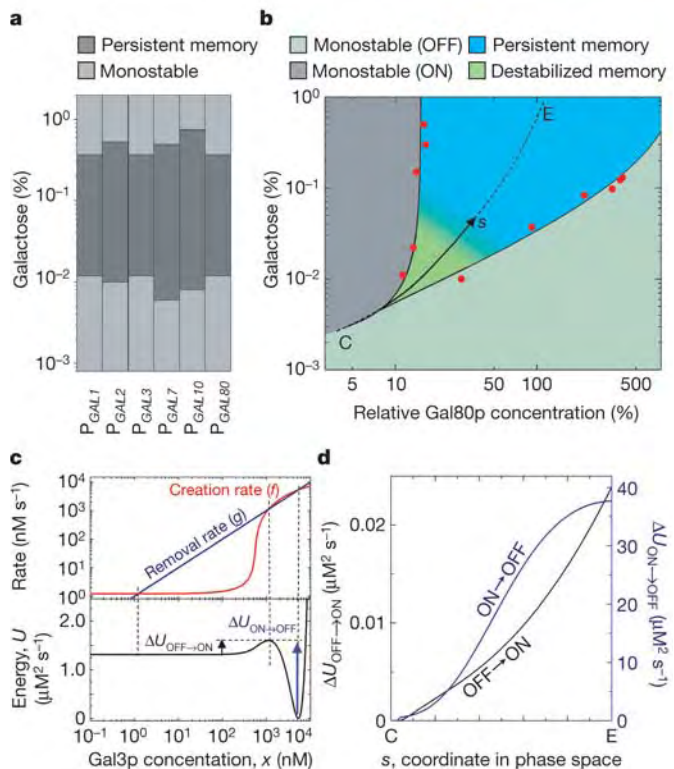


Figure 3 Stability analysis of the *GAL* network. **a**, The range over which the system displays persistent memory is similar for different strains (columns from left to right: MA0207, MA0208, MA0231, MA0212, MA0213, MA0242) in which the reporter gene is driven by different *GAL* promoters. **b**, System behaviour as a function of the control parameters: the external galactose concentration and the intracellular Gal80p concentration. *GAL80* expression is controlled by a doxycycline-inducible promoter (MA0188) and is measured relative to wild-type Gal80p expression (induced by 0.5% galactose; Supplementary Fig. S4). Red circles indicate experimentally determined boundaries between different system behaviours, and the black solid lines represent the theoretical prediction based on the regulatory network depicted in Fig. 1. The critical point is defined as C. The coordinate s defined on the path from C to the endpoint E is used in **d** to demonstrate the concept of energy barriers. **c**, The energy landscape was calculated by integrating the difference between the creation $f(x)$ and destruction rates $g(x)$ of Gal3p with respect to the Gal3p concentration denoted by x . The corresponding equation is $U(x) = -\int_0^x [f(x') - g(x')] dx'$. **d**, Magnitude of the energy barriers as a function of the coordinate s from C to E as defined in **b**.

depends strongly on the system parameters. The barriers are high in the 'persistent memory' region and vanish as the critical point C is approached at lower *GAL80* expression level (Fig. 3b, d).

These predictions were verified by measuring the switching rates between the two states (Fig. 4). For this purpose we monitored the time evolution of the expression distribution as the system approached steady state. An example of a bimodal distribution is shown in Fig. 4a (top graph) obtained from the *GAL80* loop knockout operating in the 'destabilized memory' region (Fig. 3b): some cells had high Gal4p activity but other cells had 100-fold lower Gal4p activity. From this mixed population, two subpopulations were sorted representing the two extremes of this distribution (Fig. 4a, $t = 0$). Subsequently, these two subpopulations were grown separately in the same medium as before sorting. The example in Fig. 4a shows that both subpopulations relaxed back to the same bimodal distribution. These experiments therefore rule out static disorder and indicate that cells switch back and forth between expression states. The switching rates strongly depended on the Gal80p level and the galactose concentration (Fig. 4b). When the system operated in the 'destabilized memory' region close to the critical point C (Fig. 3b) the system approached steady state in about 10 h (Fig. 4b, circles). However, if the system operated closer

to the boundary between the destabilized and persistent memory regions this process took much longer (Fig. 4b, triangles). At sufficiently high barriers, cells were almost irreversibly locked in one of the cellular states.

Figure 4c shows the experimental escape rates as a function of the calculated energy barrier. Escape rates decreased precipitously as the height of the energy barrier increases, a correlation typical of noise-induced transitions bounded by energy barriers. For the same barrier height, escape rates from the ON state were larger than escape rates from the OFF state. Indeed, the fluctuations in *GAL3* expression were larger in the ON state than in the OFF state (Supplementary Fig. S5). However, given the complexity of this network and the presence of multiple noise sources^{24,25}, it is unlikely that *GAL3* expression fluctuations can solely account for this difference in switching rates.

Next we explored whether cellular memory could affect the growth rate of a population. Increased growth rate of the ON cells relative to the OFF cells was observed when the cells were grown in medium in which galactose was the predominant carbon source (Supplementary Fig. S6). Because the fraction of ON cells in a population is strongly dependent on cellular memory, this indicates that memory might be a significant concept in understanding the fitness of a population in habitats in which carbon sources are fluctuating.

The galactose signalling pathway of budding yeast has the potential for reliably storing information on previous galactose exposures for hundreds of generations. A core positive feedback loop through *GAL3* is necessary for this cellular memory, whereas a negative feedback loop through *GAL80* competes with the positive *GAL3* loop and reduces the potential for memory storage. Consistently, when the negative feedback loop is opened and Gal80p levels are controlled constitutively, the memory persistence can be tuned from hours to months. A quantitative understanding of cellular differentiation would require a similar system-level approach to the one that we employed for the galactose signalling pathway. Systematically opening the underlying feedback loops complemented by stochastic and stability analyses provides a quantitative tool for identifying network architectures responsible for the stability of differentiated cellular states. □

Methods

Plasmid and strain constructions

KpnI-promoter-*Bam*HI, *Bam*HI-YFP-*Eco*RI fragments were cloned into pRS402 backbone upstream of *CYC1* transcriptional terminator (for *P_{GAL4}* promoter, *Bgl*II was used instead of *Bam*HI). The *P_{GAL1}*, *P_{GAL2}*, *P_{GAL3}*, *P_{GAL4}*, *P_{GAL5}*, *P_{GAL6}*, *P_{GAL7}*, *P_{GAL10}*, *P_{GAL80}* and *P_{MYO2}* promoter sequences correspond to 669, 960, 830, 633, 590, 724, 667, 632 and 677 base-pair regions upstream of the start codon of the respective genes. The *GAL3* and *GAL80* genes were cloned downstream of *KpnI*-*P_{TET2}*-*Bam*HI as *Bam*HI-*Eco*RI fragments into pRS306. *P_{GAL3}*-*GAL3*, *P_{GAL80}*-*GAL80*, *P_{TET2}*-*GAL3* and *P_{TET2}*-*GAL80* lacking the stop codon were cloned upstream of *Bam*HI-*CFP*-*Eco*RI as *KpnI*-*Bam*HI fragments in pRS306. All strains were derived from W303. Polymerase chain reaction and Southern blotting were used to verify the integrations. Complete descriptions of the strains used in this study can be found in Supplementary Information.

Growth conditions and media

Cultures were grown in synthetic dropout media with the appropriate amino-acid supplement and 2% raffinose as a carbon source. Media used for 'galactose history' experiments were supplemented with 2% galactose, whereas media for 'raffinose history' experiments contained raffinose as the sole carbon source. Cells grown overnight were diluted to a D_{600} value such that after 27 h of growth the galactose concentration in the medium would not change by more than 10% (Supplementary Information). Cells were grown at 30 °C. After the induction period of 27 h, the expression distributions were determined by flow cytometer (FACSscan; Becton Dickinson). For cell sorting, the bimodally distributed cells were sorted into OFF and ON cells as indicated in Fig. 4a.

Gal3p and Gal80p levels in inducible strains relative to wild-type levels

To determine how Gal3p and Gal80p levels in strains MA0182 and MA0188, respectively, compared with native levels of Gal3p and Gal80p in wild-type cells, strains were constructed in which the carboxy terminus of Gal3p and Gal80p was fused to CFP. The levels of doxycycline-induced *GAL3* and *GAL80* expression relative to the wild-type expression were determined by fluorescence microscopy. We found that the addition of 0.07 and 0.05 $\mu\text{g ml}^{-1}$ doxycycline resulted in the expression of wild-type levels (wild-type

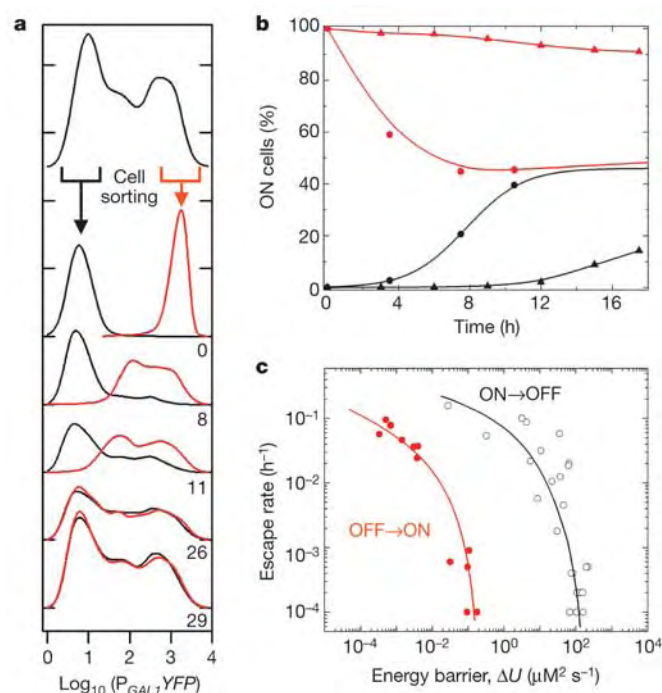


Figure 4 Stochastic switching dynamics between two stable expression states (strain MA0188). **a**, An initially bimodal population (0% galactose and $0.01 \mu\text{g ml}^{-1}$ doxycycline, 27 h of growth after raffinose history) was sorted by fluorescent activated cell sorting into cells expressing low and high levels of YFP. At different times, indicated in hours after sorting, expression distributions were measured. **b**, The fraction of ON cells as a function of time after sorting in the absence of doxycycline and galactose (filled circles) and in the presence of $0.01 \mu\text{g ml}^{-1}$ doxycycline and 0.04% galactose (filled triangles). Black and red solid lines (guides to the eye) reflect the sorted population that was initially fully OFF and ON, respectively. The standard error on the fractions is smaller than 5% as determined by reproducing a representative histogram 10 times. **c**, Experimentally measured escape rates (Supplementary Information) as a function of the calculated energy barriers. Black lines and symbols indicate experimentally determined escape rates from the ON state; red lines and symbols denote escape rates from the OFF state. Energy barriers were calculated with experimentally determined parameters obtained from fitting the experimentally determined boundaries (Fig. 3b, red circles) to the network model (Fig. 3b, solid black lines; Supplementary Information). The solid lines are guides to the eye.

cells fully induced with 0.5% galactose) of Gal3p and Gal80p, respectively (Supplementary Figs S2 and S4).

Determination of galactose consumption rate

To determine the galactose consumption rate, aliquots from cultures were filtered and the galactose concentration of the cell-free medium was analysed as follows. β -Galactose dehydrogenase was used to oxidize galactose in the presence of 2.5 mM NAD⁺ dissolved in a buffer containing 50 mM imidazole and 5 mM MgCl₂ pH 7.0 (ref. 30). Conversion of NAD⁺ into NADH was followed spectrophotometrically at 340 nm.

Received 2 February; accepted 9 March 2005; doi:10.1038/nature03524.

- Freeman, M. Feedback control of intercellular signalling in development. *Nature* **408**, 313–319 (2000).
- Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
- Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676 (2004).
- Xiong, W. & Ferrell, J. E. Jr. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* **426**, 460–465 (2003).
- Shykind, B. M. *et al.* Gene switching and the stability of odorant receptor gene choice. *Cell* **117**, 801–815 (2004).
- Markevich, N. I., Hoek, J. B. & Kholodenko, B. N. Signalling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J. Cell Biol.* **164**, 353–359 (2004).
- Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
- Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* **20**, 2528–2535 (2001).
- Atkinson, M. R., Savageau, M. A., Myers, J. T. & Ninfa, A. J. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell* **113**, 597–607 (2003).
- Isaacs, F. J., Hasty, J., Cantor, C. R. & Collins, J. J. Prediction and measurement of an autoregulatory genetic module. *Proc. Natl Acad. Sci. USA* **100**, 7714–7719 (2003).
- Angeli, D., Ferrell, J. E. Jr. & Sontag, E. D. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc. Natl Acad. Sci. USA* **101**, 1822–1827 (2004).
- Lai, K., Robertson, M. J. & Schaffer, D. V. The sonic hedgehog signaling system as a bistable genetic switch. *Biophys. J.* **86**, 2748–2757 (2004).
- Biggar, S. R. & Crabtree, G. R. Cell signaling can direct either binary or graded transcriptional responses. *EMBO J.* **20**, 3167–3176 (2001).
- Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
- Suzuki-Fujimoto, T. *et al.* Analysis of the galactose signal transduction pathway in *Saccharomyces cerevisiae*: interaction between Gal3p and Gal80p. *Mol. Cell. Biol.* **16**, 2504–2508 (1996).
- Timson, D. J., Ross, H. C. & Reece, R. J. Gal3p and Gal1p interact with the transcriptional repressor Gal80p to form a complex of 1:1 stoichiometry. *Biochem. J.* **363**, 515–520 (2002).
- Peng, G. & Hopper, J. E. Gene activation by interaction of an inhibitor with a cytoplasmic signaling protein. *Proc. Natl Acad. Sci. USA* **99**, 8548–8553 (2002).
- Mizutani, A. & Tanaka, M. Regions of GAL4 critical for binding to a promoter *in vivo* revealed by a visual DNA-binding analysis. *EMBO J.* **22**, 2178–2187 (2003).
- Johnston, M., Flick, J. S. & Pexton, T. Multiple mechanisms provide rapid and stringent glucose repression of GAL gene expression in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **14**, 3834–3841 (1994).
- Wieczorke, R. *et al.* Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett.* **464**, 123–128 (1999).
- Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & van Oudenaarden, A. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427**, 737–740 (2004).
- Bhat, P. J. & Hopper, J. E. The mechanism of inducer formation in *gal3* mutants of the yeast galactose system is independent of normal galactose metabolism and mitochondrial respiratory function. *Genetics* **128**, 233–239 (1991).
- Rohde, J. R., Trinh, J. & Sadowski, I. Multiple signals regulate GAL transcription in yeast. *Mol. Cell. Biol.* **20**, 3880–3886 (2000).

- Blake, W. J., Kaern, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
- Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
- Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. Noise-based switches and amplifiers for gene expression. *Proc. Natl Acad. Sci. USA* **97**, 2075–2080 (2000).
- Bialek, W. Stability and noise in biochemical switches. *Adv. Neural Info. Process.* **13**, 103–109 (2001).
- Kepler, T. B. & Elston, T. C. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* **81**, 3116–3136 (2001).
- Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304 (1940).
- Sillje, H. H. *et al.* Effects of different carbon fluxes on G1 phase duration, cyclin expression, and reserve carbohydrate metabolism in *Saccharomyces cerevisiae*. *J. Bacteriol.* **179**, 6560–6565 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Thattai for stimulating discussions and helpful suggestions on measuring the escape rates; J. Pedraza, M. Kardar, O. Ozcan and S. Cabi for useful discussions; B. Kaufmann for help with image analysis; and B. Pando for comments on the manuscript. M.A. acknowledges support by an MIT Presidential Fellowship endowed by Praecis Pharmaceuticals, Inc. A.B. is a Long Term Fellow of the Human Frontiers Science Program. This work was supported by a NSF-CAREER and a NIH grant.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.v.O. (avano@mit.edu).

corrigendum

Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic

Matthew M. Mills, Celine Ridame, Margaret Davey, Julie La Roche & Richard J. Geider

Nature **429**, 292–294 (2004).

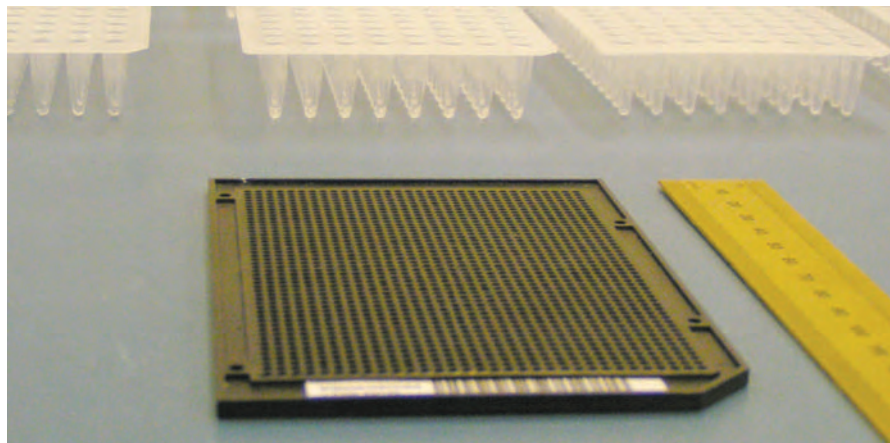
In this Letter to *Nature*, the chlorophyll *a* data presented in Fig. 1d–f and in Table S1 of the Supplementary Information are an order of magnitude too low owing to a calculation mistake. This error does not alter the conclusions of our paper. □

Replicating success

PCR often gets taken for granted, but there are ways of making it faster, more accurate and easier to perform. Pete Moore investigates.

As a means of rapidly copying a selected template sequence from a DNA mixture *in vitro*, PCR by itself and in combination with other techniques has found a vast range of applications. These range from sequence detection and isolation for research, forensics and species identification to detecting mutations and polymorphisms and amplifying RNA-derived cDNAs for microarray analysis of gene expression. As well as standard PCR, the technique now comes in the form of real-time quantitative PCR (real-time PCR or qPCR). This uses fluorescent probes to monitor the amount of product at the end of each cycle, and real-time PCR machines look for the cycle at which they can first detect fluorescence. This relates to the number of copies of original template — the greater the number of starting copies, the fewer cycles are needed to reach fluorescence detection.

PCR can also be used to monitor RNA by adding a reverse transcriptase enzyme at the beginning to generate a DNA template. This reverse transcription PCR (RT-PCR) can then be taken a step further by adding the quantification protocols, resulting in real time RT-PCR.



More reactions in the same space with 1,536 wells.

There can be few life-science laboratories without a PCR thermal cycler tucked in a corner, happily churning out short DNA sequences to order with tried and tested protocols. But newer applications for PCR, such as single-nucleotide polymorphism (SNP) detection and screening, need faster throughput, and this is now achievable.

One approach is to abandon the traditional 96-well plate in favour of 384 wells or

more. The new high-throughput 7900HT fast real-time PCR system from Applied Biosystems in Foster City, California, takes 96- and 384-well plates and runs a full set of amplification cycles in about 35 minutes. “Our new high-speed system can alleviate some of the burden of instrument sharing by reducing cycling time,” says Peter Dansky, senior director and general manager of core PCR at Applied Biosystems. If you need

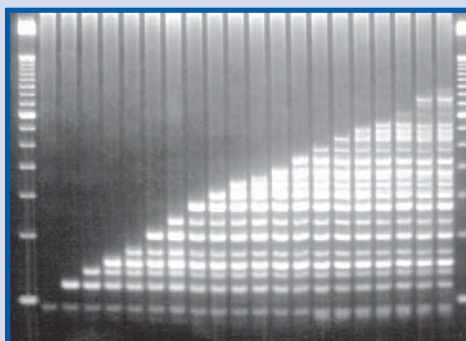
HOT FROM THE VENT

Isolated from the heat-loving bacterium *Thermus aquaticus*, Taq DNA polymerase launched PCR. It was great, and versions of the original recombinant Taq are still widely used — but its error rate of between 1 in 10,000 and 1 in 50,000 base pairs (bp) is too high for some applications, such as the detection of single-nucleotide mutations. You can now take your pick from a plethora of Taq-based polymerases engineered to have higher fidelity and to go faster. And if damaged DNA is the problem, Restorase from Sigma-Aldrich of St Louis, Missouri, is a mixture of Sigma's AccuTaq blend and a repair enzyme, and works with fragment lengths from 200 to 20,000 bp.

Developers have also gone back to the planet's hot springs and hydrothermal vents to find a new generation of thermostable polymerases with the 3'–5' exonuclease proofreading capacity that makes for higher accuracy. The archaeal genus *Pyrococcus* has been mined for high-fidelity DNA polymerases with accuracies some 40 to 50 times greater than Taq. *P. abyssi* is the source of the Isis proofreading DNA polymerase from Qbiogene of Irvine, California, with an error rate of one mismatched base per 1.5 million bases per duplication. Other hot offerings come from Stratagene, of La Jolla, California, which claims an error rate of 1 in around 660,000 for its *Pfu*

DNA polymerase from *P. furiosus*, while Bio-Rad of Hercules, California, has the highly processive iProof High-Fidelity DNA polymerase, a *Pyrococcus*-type polymerase fused to a double-stranded DNA-binding protein to give additional grip, which the company claims is 50 times more accurate than Taq. As well as a DNA polymerase isolated from *P. woelsi*, Roche Applied Science of Indianapolis, Indiana, offers a thermostable reverse transcriptase from *Carboxydotherrhus hydrogenoformans* for RT-PCR, and Toyobo Company of Osaka, Japan, supplies a very fast DNA polymerase from *P. kodakaraensis* (now renamed *Thermococcus*).

Proofreading enzymes are generally more finicky than Taq or enzyme blends. To address this problem, Invitrogen's AccuPrime Pfx DNA polymerase is a *P. kodakaraensis* polymerase in a mix containing the company's proprietary AccuPrime accessory proteins (also available as a mix with the company's Taq polymerase), which improve specificity by ensuring that primers only bind to their complementary sequence. The QuantiTect Multiplex PCR kits from Qiagen, of Hilden, Germany, also aim to provide trouble-free multiplex PCR by including a novel reaction chemistry in the buffer that helps avoid competition between PCR products and ensures efficient primer hybridization.



Clean multiplexing with AccuPrime.



The PCRJet from MegaBase gives fast PCR with large samples.

“Once you have got to four colours you have four sets of primers and four probes, and you have to stop these cross-reacting with each other and forming primer dimers,” says Chris Helps from the School of Clinical Veterinary Science at the University of Bristol. “When you increase the number of targets the reactions become very complex — you also need spectrally distinct dyes to minimize cross-talk between channels.”

Blowing hot and cold

How about really cutting the time down? The RapidCycler thermal cycler from Idaho Technology of Salt Lake City, Utah, blows blasts of hot and cold air through the reaction chamber, which gives near instantaneous temperature changes and rapid heat exchange with the samples. The RapidCycler 2 will do a 30-cycle run in 15 minutes, carrying 48 samples in either glass microcapillary tubes or the standard 1.5-mm reaction cuvettes that will also fit the widely used LightCycler, available from Roche Applied Science in Indianapolis, Indiana.

And it is possible to go faster still. The PCRJet thermocycler developed by a multidisciplinary team under the brand name MegaBase Research Products, in Lincoln, Nebraska, drives a mixture of hot and cold gas through the reaction chamber at 45 miles per hour. “The velocity of the air stream is so high that we are definitely in the turbulent region, which ensures that the heat transfer to the sample-containing capillary tubes is maximal. We can do 30 cycles of PCR with

more than 384 wells, more manufacturers are now making 1,536-well plates, including Corning, of Corning, New York, evotec technologies of Hamburg, Germany, and KBiosciences of Hoddesdon, UK. The problem with the larger plates is getting robotic support, but Victor Crew, sales manager at KBiosciences, says that it is possible to modify automated pipettors like the Plate-Mate from Matrix Technologies in Beverly, Massachusetts, to use them. The Equator HTS and Latitude pipetting systems from Deerac Fluidics in Dublin, Ireland, will also take 1,536-well plates and will fill one in less than 15 seconds.

Throughput can also be increased by multiplexing — running more than one spe-

cific amplification reaction in a single tube. One approach is to use different colours of fluorogenic dyes to detect the different products. The new Mx3005P real-time PCR system from Stratagene of La Jolla, California, allows five different fluorogenic dyes to be used simultaneously. Using the company's FullVelocity probe-based real-time reagents, the machine will complete a 40-cycle two-step real-time PCR reaction in about 50 minutes. Cepheid of Sunnyvale, California, make a real-time thermal cycler with four-colour optics and 96 independently programmable reaction holders that claims to get the job done in as little as 20 minutes.

There is probably a limit to how many dyes can be added to a single reaction tube.

AMPLIFYING THE SIGNAL

When target amplification is not needed, northern blotting and other methods based on nucleic acid hybridization are convenient for detecting and measuring specific RNAs. Such methods include the Quantikine range for measuring specific cytokine mRNAs from R&D Systems of Minneapolis, Minnesota. The problem with direct hybridization is generating a strong signal from a few target molecules.

Chad Mirkin, director of the Northwestern University Institute for Nanotechnology, at Evanston, Illinois, thinks he has cracked the problem.



Chad Mirkin is barcoding proteins.

He is co-founder of Nanosphere in Northbrook, Illinois, which produces BioBarcode, a sensitive detection system that can detect DNA sequences, but is proving to be just as good at spotting specific proteins. “PCR is amazing because it takes what you are trying to detect and duplicates a portion of it so that you have enough

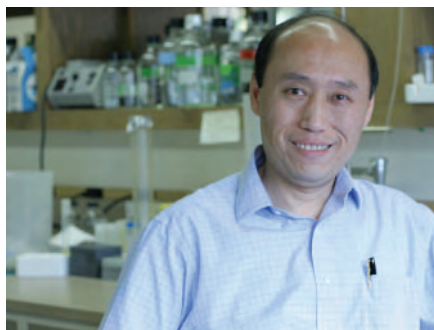
to detect. If you could do that for proteins, I would argue that you could have as big or a bigger impact, especially if you could do it without needing to use enzymes,” he says.

Mirkin hopes to make that impact with BioBarcode, which works as a two-component immunological sandwich assay with some novel twists. Antibody molecules specific for one site on the target protein are tagged with 13-nm diameter gold nanoparticles, while larger paramagnetic particles are coated in a second antibody that recognizes a separate site on the target. If present, the target protein will grab both probes, and the complexes can be isolated by magnetic separation.

So far, not too unconventional. In Mirkin's system, however, the gold nanoparticle is smothered in hundreds or thousands of identical oligonucleotides, each strand hybridized to a short sequence that acts as a molecular barcode. Nanoparticles with different antibodies carry different barcodes, allowing multiplexing. After separation, the DNA barcodes are released and their identity determined by scanning on a biochip. The initial signal is thus amplified hundreds to thousands of times, allowing as few as 10 molecules of protein to be detected.

Randy Lewis at the University of Wyoming is using BioBarcodes to search for prions associated with the elk version of mad cow disease. “BioBarcodes give us an assay that is 1,000 times more sensitive than anything else,” he says. “We should now be able to detect early onset and environmental contamination at lower levels than could be done before.”

P.M.



Huimin Kong: using helicase avoids having to put the heat on.

amplicons of anything from 100 to 600 base pairs in 2 to 3 minutes," says Hendrik Viljoen of the department of chemical engineering at the University of Nebraska in Lincoln, one of the designers. "Going faster than 5 minutes doesn't really gain much for the working scientist, since it usually takes longer than that to mix the PCR reagents," says team member Michael Nelson, "so we have backed off on speed and are now primarily concerned with system engineering for reliability and ease of use."

The PCRJet takes eight samples of 20–100 µl at a time. "Talking to people in industry, we have found that in areas like infectious disease detection there is strong resistance to too small volumes. You need a big enough lump of sample because in the early stages of disease development there may be very few organisms present in a sample," says Viljoen. PCRJet needs a fast enzyme, and it uses

KOD Pol from *Pyrococcus kodakaraensis* from Toyobo Company of Osaka, Japan. Toyobo's Hideki Hayami, who is collaborating with MegaBase, says this can copy DNA at a rate of around 300 nucleotides per second.

Getting personal, Stratagene's Mx3000P is a four-colour optics real-time PCR machine for personal and small lab use, while the 46-well MJ Mini thermal cycler from Bio-Rad of Hercules, California (which recently acquired the manufacturers MJ Research) can be upgraded to a two-colour real-time machine with the retrofit of a MiniOpticon detection system.

Isothermal alternatives

Given that the physics of heating and cooling a PCR system can be problematic, a few pioneers are developing isothermal procedures — PCR at a uniform temperature — with an eye mainly on the clinical diagnostic market.

One approach is known as the ramification amplifying method (RAM). Invented by David Zhang of Mount Sinai School of Medicine in New York in 1994, RAM is licensed to Hamilton Thorne Biosciences of Beverly, Massachusetts, by Mount Sinai, which was recently granted an additional US patent on the method.

Target DNA is first isolated by capture probes linked to magnetic beads. A given target DNA is then detected by RAM, which employs a single-stranded DNA 'C-probe' that contains 3' and 5' sequences complementary to the target. If the C-probe hybridizes accurately with the target DNA,

both ends of the probe bind close together and are joined by a ligase to form a circle. Once the circle is formed, this binds a primer at an internal site, which is extended by a strand-displacing DNA polymerase (for example, phi29 or Bst). As the polymerase travels around the circle it displaces the strand created on previous circuits, creating a long chain, which can itself be duplicated by other primers and enzymes.

As the displaced DNAs are single stranded, primers can bind at a consistent temperature, removing the need for any thermocycling during amplification. However, "circle formation does require annealing of the C-probe to the targets and for long double-stranded DNA targets, a denaturation step would be beneficial," explains David Lane, vice-president of research and development for Hamilton Thorne Biosciences.

Binding is highly specific, making it a useful tool for SNP detection. "Since RAM uses a universal primer, there is also no need for primer balancing, making high-level multiplexing as straightforward as adding multiple C-probes to an assay," says Lane.

A completely different approach is that of Huimin Kong and colleagues, who came up with the idea of using a DNA helicase to separate the DNA strands rather than heat while working at New England Biolabs in Beverly, Massachusetts. "While other so-called isothermal techniques need an initial 95 °C DNA denaturation step, helicase-dependent amplification (HDA) can be performed in true isothermal conditions," Kong says. Once

PHOTOCOPIERS FOR DNA

Rather than amplifying just the DNA between two primers, many researchers want to amplify entire genomes, to make archive copies of the total DNA from a unique biopsy sample, for example. "If you have a piece of paper with valuable information on it, obviously you don't want to lose the paper, but instead photocopy it many times, store some copies appropriately and use others," says Andy Betera vice-president of product management of GE Healthcare, based in Little Chalfont, UK. Whole-genome amplification (WGA) systems aim to act as photocopiers for DNA. "From the point of

view of the molecular epidemiologist or molecular geneticist, the overwhelming potential advantage of the newer methods of WGA is the possibility of producing additional genomic DNA, or amplified WGA DNA, for genotyping, sequencing or other types of genetic analysis like loss of heterozygosity (LOH)," says Andrew Bergen, staff scientist at the National Cancer Institute in Bethesda, Maryland.

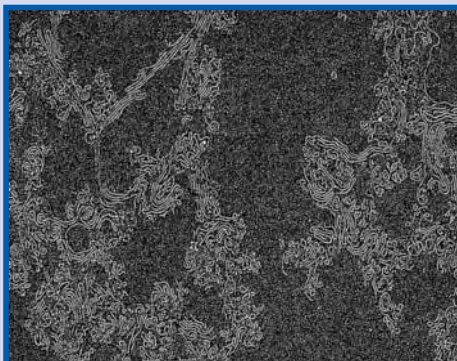
"For the past 12 years there have been PCR-based methods of WGA with about a 10 or 15% locus bias, meaning that 10 or 15% of the genetic loci in the genome would not be amplified," explains Bergen. Newer methods can avoid this loss in the right conditions. But problems really arise when you have only a very small sample from a slide or biopsy to play with. Bergen says

colleagues who have almost exhausted the material from a particular sample ask him: "Can this sample be rescued?" And no WGA method is really effective with less than 10 ng of genomic DNA, he says.

On the commercial front, GE Healthcare's GenomiPhi DNA amplification kit uses the highly processive single-strand displacing DNA polymerase from phage phi29. Short primers are bound randomly throughout the genome and the enzyme copies the DNA, starting from each primer and displacing the primer ahead of it. The result is a soup of genomic fragments

of varying lengths, averaging around 10,000 bases. The phi29 method is able to efficiently amplify a whole genome with no loss of sequence if high molecular weight DNA is used as the starting material.

The GenomePlex WGA kit from Sigma-Aldrich uses technology licensed from Rubicon Genomics of Ann Arbor, Michigan. In this method, the genomic DNA is initially broken up into fragments of around 400 bases long, which are then attached to an identical sequence. The collection is amplified by PCR using a primer that recognizes this sequence. Because short fragments are being copied, this method can cope with starting DNA of high or low molecular weight.



GenomiPhi-amplified salmon sperm.

the helicases have unwound the target, sequence-specific primers can bind to the single-stranded DNA and be extended, as in standard PCR.

Along with former colleagues, Kong has founded BioHelix Corporation in Beverly, with the aim of commercializing HDA. BioHelix's first commercial product is a teaching kit, marketed by the Carolina Biological Supply Company in Burlington, North Carolina, for students to carry out molecular diagnosis of sickle-cell anaemia in the classroom, without the need for an expensive thermocycler. The IsoAmp tHDA DNA amplification kit for research is due to launch this month.

Primers and probes made easy

For real-time PCR, the combinations of specific primers and the fluorescent oligonucleotide probes that detect template amplification are key. Researchers who hate designing probes and primers can now turn to online databases. PrimerBank, developed by Xiaowei Wang of the department of molecular biology at Harvard University and Massachusetts General Hospital, Boston, contains over 300,000 predicted primers for human and mouse genes generated computationally by a design algorithm. They have tested over 1,000 primer pairs and found a design success greater than 99% as defined by single PCR products and reasonable amplification efficiency, says Wang.

The Quantitative PCR Primer Database (QPPD), coordinated at the National Cancer Institute in Bethesda, Maryland, provides

information about published primers and probes for quantitating human and mouse mRNA by RT-PCR.

Another source of published primer-probe sets for real-time PCR, with 3,376 entries so far, is RTprimerDB run by Jo Vandesompele and Filip Pattyn at the University of Ghent, Belgium, which can be searched by gene name, oligonucleotide sequence or NCBI's EntrezGene or SNP ID. "If people find the primer-probe pair they want they can click through to the PubMed identifier, and it also contains the details of the person who submitted the sequence," says Vandesompele. He hopes that coordinating the primer-probe sets that people use should increase standardization between labs, making it easier to compare results. Vandesompele also sees perils in the common use of a single housekeeper gene to normalize results in gene-expression studies, and his free geNorm applet for Excel will determine how many housekeeper genes you need for an accurate analysis. But he laments the fact that most analysis software only allows one housekeeper gene to be entered.

First described by Jesper Wengel in 1998, locked nucleic acids (LNA) are slowly making their way on to the PCR scene for use as probes. Their key advantage is their restricted conformational flexibility, which gives them great thermal stability and excellent mismatch discrimination when complexed with complementary DNA or RNA. At the Charité University Medical Centre in Berlin, Germany, Oliver Goldenberg and Lutz Hamann

are using LNAs to quantify the species-specific 16S rDNA from multiple bacteria in a single PCR reaction. Their interest is in the intestinal flora that keep us healthy and they want a fast method of detecting the bacteria present without having to cultivate them. With a standard reporter dye such as SYBR Green, which fluoresces when it binds double-stranded DNA, only one specific feature can be assayed per sample — either a total bacterial count or identification of a single species. The widely used fluorescent resonance emission transfer (FRET)-based probes, such as the TaqMan system from Applied Biosystems of Foster City, California, aren't suitable either, as they require a recognition sequence of some 40 bases and "you will hardly find such long conserved regions in 16S rDNA," says Goldenberg. "But the higher melting point of LNAs means that specificity can be achieved with shorter sequences."

Designing LNA probes can pose problems (see 'Simplifying the probe set'). "We have been testing LNA probes for SNP detection and have found the design parameters to be significantly different to standard TaqMan probes," says Helps. He thinks the problem lies in the probe-design software: the most commonly used design software was written for the established fluorogenic probes but doesn't work as well for LNAs.

PCR has its twentieth birthday this year and has stood the test of time. Like the DNA it analyses it is evolving, and the next 20 years should be equally exciting. ■

Pete Moore is a freelance writer based near Bristol, UK.

SIMPLIFYING THE PROBE SET

The conventional wisdom on probe-based real-time PCR assays says that if you have 10,000 genes to detect you will need 10,000 different probes to detect the PCR products. And to provide the required specificity, the probes need to be at least 18 nucleotides long. A team of researchers and bioinformaticians at Exiqon in Vedbaek, Denmark, is challenging this dogma. By scanning the genomes of key species they have come up with a set of 90 locked nucleic acid (LNA) 8- and 9-mer probes that they claim can be used to identify every gene on which you will ever work. Because of the greater

rigidity of LNAs, even a single mismatch with the potential target will seriously affect binding and prevent generation of signal, thus giving the same specificity as a longer regular oligonucleotide probe.

The probes come to life when used with Exiqon's software, which for each target gene designs an optimum primer pair in combination with a

specific probe. "The short LNA probes enable the technology, but the software is the key," says Exiqon's new technology-development manager, Peter Mouritzen. To identify the most specific combination, the software performs what Exiqon calls *in silico* PCR. This checks primer-probe predictions against the entire genome and transcriptome of the organism, minimizing the risk of possible mispriming which could produce a false-positive signal from a real-time PCR assay.

For RT-PCR you want to look at the transcribed sequence and not the genomic DNA, which could contaminate the sample, so the software increases specificity by choosing a section where the primers span an exon-exon splice junction. While the primer could hit the genomic DNA, it will not generate a signal, because the intron is likely to make the product so long that it will not be amplified efficiently.

Users buy the LNA probes from Exiqon but can get their primers from any supplier. The assay design software is free online. "Normally people spend hours designing a PCR assay, and this thing does it in seconds," comments Mouritzen.

Roy Bicknell at the Weatherall Institute of Molecular Medicine at the University of Oxford, UK, agrees. "We look for new genes expressed on tumour vasculature as anticancer targets, so when we identify them by bioinformatics we have to do a lot of validation. Because we are looking at many genes we need to keep making new primers — and that is the beauty of the Exiqon system — it's for people who want to look at lots of different genes," he says.

P.M.

360DEGREES



Peter Mouritzen has designs on LNA.

Company	Products/activity	Location	URL	
PCR enzymes and kits				
Abgene	PCR reagents, quantitative PCR, microplates	Epsom, UK	www.abgene.com	
Active Motif	Reverse transcriptase, peptide nucleic acids for gene silencing, kits and reagents for cell fractionation, nucleic acid isolation	Carlsbad, California	www.activemotif.com	●
Ambion	Kits for PCR, kits and products for RNA synthesis, isolation, quantitation and analysis	Austin, Texas	www.ambion.com	
Bio-Rad	Enzymes and kits for PCR, MJ Research thermal cyclers, BaseStation DNA fragment sequencer and genotyper, labware, reagents for molecular biology	Hercules, California	www.bio-rad.com	●
Chemicon	PCR kits, molecular biology reagents and kits to analyze DNA and RNA expression in viral, prokaryotic, and eukaryotic experimental systems	Temecula, California	www.chemicon.com	
Clontech	Taq polymerase and enzyme mixes for PCR, RNAi expression kits, tools for protein purification and analysis, cellular analysis, nucleic acid purification	Mountain View, California	www.clontech.com	
EPICENTRE	Enzymes for PCR and RT-PCR, products for transposon-based genetic analysis, DNA and RNA purification	Madison, Wisconsin	www.epicentre.com	
Finnzymes	Enzymes, PCR kits and transposon products for genomics and proteomics	Espoo, Finland	www.finnzymes.fi	●
GE Healthcare	Instruments, equipment and products for genomics, proteomics and cellular assays whole-genome amplification system (GenomiPhi), enzymes for PCR	Little Chalfont, UK	www.amersham.co.uk	●
Genetix	HySpot - Probe Dye-Labeling Kit, robots for genomics and proteomics	New Milton, UK	www.genetix.co.uk	
GenScript	Kits and services for cell biology and molecular biology, PlantDirect, TissueDirect, and BloodReady Multiplex PCR systems	Piscataway, New Jersey	www.genscript.com	●
Invitrogen	Enzymes and kits for PCR, cloning, genomics, proteomics, molecular and cell biology, genotyping kits, antibodies	Carlsbad, California	www.invitrogen.com	●
New England Biolabs	Reagents and kits for molecular biology, DyNAmo qPCR and qRT-PCR kits	Beverly, Massachusetts	www.neb.com	
Novagen	Enzymes, reagents and kits for PCR and RT-PCR, PCR kits with KOD high-fidelity DNA polymerase	Madison, Wisconsin	www.novagen.com	
Qbiogene	Enzymes and kits for PCR	Irvine, California	www.qbiogene.com	
QIAGEN	Enzymes and kits for PCR, QuantiTect Multiplex PCR Kits, QuantiTect Primer Assays for SYBR Green detection, RNeasy for RNA purification, reagents, kits and instruments for genomics, proteomics, molecular and cell biology	Valencia, California	www.qiagen.com	
Roche Applied Science	Enzymes and kits for PCR, LightCycler thermal cycler for real-time PCR, reagents and kits for molecular biology, functional genomics and proteomics research	Lewes, UK	www.roche-applied-science.com	
Sigma-Aldrich	Reagents, antibodies, and biochemicals for life-sciences research, Genomeplex whole-genome amplification system	St Louis, Missouri	sigma-aldrich.com	
Stratagene	Enzymes and kits for PCR, Mx3000P and Mx4000 real-time PCR systems	La Jolla, California	www.stratagene.com	
TaKaRa Bio	Reagents, kits and custom services for molecular biology, genomics, proteomics proteomics research, DNA microarrays, protein refolding kit, Intelligene biochip	Shiga, Japan	www.takara-bio.co.jp/english	
Toyobo	Enzymes for PCR,	Osaka, Japan	www.toyobo.co.jp	
Transgenomics	Enzymes for PCR and mutation detection, genotyping	Omaha, Nebraska	www.transgenomic.com	
Wako Chemicals	Specialty chemicals, bioproducts and clinical diagnostic reagents	Richmond, Virginia	www.wakousa.com	

Thermal cyclers and PCR systems

Applied Biosystems	Real-time PCR systems, TaqMan probes for real-time PCR	Foster City, California	home.appliedbiosystems.com	●
Beckman Coulter	Automated tools for molecular biology, biochemistry, genomics and proteomics	Fullerton, California	www.beckmancoulter.com	
Bioneer	Custom oligonucleotides, enzymes for PCR, thermal blocks	Daejeon, Korea	www.bioneer.com	
Brinkmann	Laboratory instrument suppliers, software, consumables	Westbury, New York	www.brinkmann.com	
Cepheid	Instruments and equipment for DNA analysis and PCR	Sunnyvale, California	www.cepheid.com	
Corbett Research	Rotor-Gene 3000 thermal cycler for real-time PCR, equipment for DNA analysis	Sydney, Australia	www.corbettresearch.com	
Idaho Technology	RapidCycler 2 hot-air thermal cycler	Salt Lake City, Utah	www.idahotech.com	
Megabase Research Products	PCRJet thermal cycler, DNA modification enzymes, immunochemical reagents	Lincoln, Nebraska	www.pcrjet.com	
Tecan	Genesis and Freedom EVO workstations and automated solutions for genomics and proteomics	Männedorf, Switzerland	www.tecan.com	
Technne	Laboratory equipment, thermal cyclers	Burlington, New Jersey	www.technne.com	
Thermo Electron	PCR Sprint and MBS Satellite thermal cyclers, instruments and reagents for molecular biology, custom peptide and oligonucleotide synthesis	Ashford, Middlesex	www.thermo.com	

Primers and probes

Dharmacon	RNA technologies, siRNA kits and arrays, custom RNAs, large-scale RNA synthesis	Lafayette, Colorado	www.dharmacon.com	
Epoch Biosciences	Custom oligonucleotides for PCR and microarrays; MGB Eclipse probe systems for gene-expression quantitation, SNP detection, and DNA detection	Bothell, Washington	www.epochbio.com	
Exiqon	Locked nucleic acid (LNA) oligonucleotides; ProbelLibrary gene-expression kits for real-time PCR validation of target genes and DNA microarray experiments	Vedbaek, Denmark	www.exiqon.com	
Integrated DNA Technologies	Custom oligonucleotides including locked nucleic acids (LNAs)	Coralville, Iowa	www.idtdna.com	●
MWG Biotech	Custom oligonucleotides, probes for quantitative PCR, microarrays	Ebersberg, Germany	www.mwg-biotech.com	
OligoEngine	Custom oligonucleotides	Seattle, Washington	www.oligoengine.com	
Prologo Primers & Probes	Custom and specialty synthesis of DNA and RNA oligos, locked nucleic acids (LNAs), probes, siRNA oligos for RNAi	Boulder, Colorado	www.prologo.com	
R&D Systems	Primer pairs for specific applications, cDNA expression kits, Quantikine mRNA quantitation kits, cytokines, antibodies, ELISA, flow cytometry kits	Minneapolis, Minnesota	www.RnDSystems.com	
RNA-TEC	Custom oligonucleotide synthesis	Leuven, Belgium	www.rna-tec.com	
Sigma-Genosys	Custom oligonucleotide primers and probes for PCR	The Woodlands, Texas	www.sigma-genosys.com	

Company	Products/activity	Location	URL
General			
Affymetrix	RNA amplification kits, microarrays	Santa Clara, California	www.affymetrix.com
Agencourt Bioscience Corporation	Products and instruments for DNA purification, genomics services: DNA sequencing, whole-genome sequencing, SAGE, SNP discovery	Beverly, Massachusetts	www.agencourt.com
Agilent	2100 Bioanalyzer for genomics and proteomics, instruments for genomics and proteomics, RNA amplification kits	Palo Alto, California	www.agilent.com
Apogent Technologies	Labware and equipment for the life sciences	Portsmouth, New Hampshire	www.apogent.com
BD Biosciences	Culture media, radioassay kits, FACS range of flow cytometers, reagents and biochemicals for molecular biology	Franklin Lakes, New Jersey	www.bd.com
Carolina Biological Supply Company	Support materials for science education	Burlington, North Carolina	www.carolina.com
Corning	Labware and laboratory equipment for the life sciences, coated slides	Corning, New York	www.corning.com
Deerac Fluidics	Equator and Latitude ranges of liquid-handling systems	Dublin, Eire	www.deerac.com
Elchrom Scientific	Tools for protein research, gel electrophoresis, membrane for Western blots, resins	Cham, Switzerland	www.elchrom.com
Eppendorf	Laboratory instrumentation and consumables for molecular and cell biology	Hamburg, Germany	www.eppendorf.com
evotec technologies	Hardware, software and bioware for measurement, miniaturization and automation	Hamburg, Germany	www.evotec-technologies.com
Gibco	Cell culture media and consumables	Carlsbad, California	www.invitrogen.com
Greiner Bio-One	Consumables, microplates, HTAplatform microplates for high-throughput arrays	Frickenhäusen, Germany	www.gbo.com/bioscience
Hamilton Company	MicroLab automated liquid-handling workstations	Reno, Nevada	hamiltoncomp.com
Matrix Technologies	Tango liquid-handling system, Platemate plate-handling system, microplates and consumables, arraying equipment	Hudson, New Hampshire	www.matrixtechcorp.com
New England Biolabs	Reagents and kits for molecular biology, HiScribe RNAi transcription kit	Beverly, Massachusetts	www.neb.com
Open Biosystems	Clones, RNAi, chromosome paints, antibody services; mouse shRNA libraries	Huntsville, Alabama	www.openbiosystems.com
OriGene Technologies	Human full-length cDNA clones, including complete GPCR, protein kinase, and nuclear hormone receptor full-length cDNA collections	Rockville, Maryland	www.origene.com
PerkinElmer Life Sciences	MultiPROBE II nucleic acid workstation, equipment and probes for PCR	Boston, Massachusetts	www.perkinelmer.com
Pierce Biotechnology	Components and consumables for automated systems in molecular biology, protein chemistry, sample handling, immunology and chromatography	Rockford, Illinois	www.piercenet.com
PreAnalytiX	Kits for nucleic acid isolation from blood samples	Hombrechtikon, Switzerland	www.preanalytix.com
Premier Biosoft	Desktop bioinformatics packages for sequence analysis, primer design (Array Designer), and two-hybrid protein interactions, SimVector 3	Palo Alto, California	www.premierbiosoft.com
Promega	Vectors, reagents and kits for genomics, proteomics and cellular analysis	Madison, Wisconsin	www.promega.com
Robbins Scientific	Equipment and disposables for molecular biology, including PCR	Sunnyvale, California	www.robsci.com
Sigma-Aldrich	Reagents and biochemicals for life-sciences research, antibodies	St Louis, Missouri	www.sigmaaldrich.com
System Biosciences	Tools for systems biology, functional genomics, proteomics, drug target discovery and validation, signal transduction pathways	Mountain View, California	www.systembio.com
Thermolyne	Hot plates, mixers and stirrers	Dubuque, Iowa	www.barnsteadintl.com
USB	Biochemicals and reagents for molecular biology, including PCR, manual DNA sequencing, mutation detection	Cleveland, Ohio	www.usbweb.com
VWR International	Laboratory product suppliers	West Chester, Pennsylvania	www.vwr.com
Services			
KBiosciences	1,536-well plates, genomics services	Hoddesdon, UK	www.kbioscience.co.uk
Capital Genomix	GeneSystem 320 differential gene-expression profiling, custom antibodies, Affymetrix microarray services	Chantilly, Virginia	www.capitalgenomix.com
Global Genomics	PCR and capillary gel electrophoresis-based 'tangerine' gene expression profiling for drug discovery	Stockholm, Sweden	www.globalgenomics.com
deCODE Genetics	Pharmacogenomics services, high-throughput PCR-based microsatellite genotyping	Reykjavik, Iceland	www.decode.com
Sequenom	MassArray PCR-based system for high-throughput genotyping, targeted SNP discovery, allele frequency analysis	San Diego, California	www.sequenom.com
ParAllele Bioscience	SNP genotyping, genome variation scanning, gene-expression analysis	San Francisco, California	www.p-gene.com
Third Wave Technologies	Invader RNA assays and genotyping kits for a wide variety of human genes	Madison, Wisconsin	www.twt.com
Althea Technologies	Real-time QPCR and QRT-PCR services and assay development	San Diego, California	www.altheatech.com
Other nucleic acid detection and amplification			
BioHelix	Helicase-dependent amplification for nucleic acid analysis	Beverly, Massachusetts	www.biohelix.com
Genisphere	SenseAmp and SenseAmp Plus kits for RNA amplification	Hatfield, Pennsylvania	www.genisphere.com
Genospectra	QuantiGene proprietary system for RNA detection and quantitation	Fremont, California	www.genospectra.com
Hamilton Thorne Biosciences	Nucleic acid detection technology, rolling circle isothermal DNA amplification technology	Beverly, MA	www.hamiltonthorne.com
Nanosphere	Verigene and Bio-barcode nanoparticle/magnetic bead-based assays for nucleic acids and proteins	Northbrook, Illinois	www.nanosphere-inc.com
NuGEN Technologies	Ovation RNA amplification process	San Carlos, California	www.nugentechnologies.com
US Genomics	Trilogy platform for detection of single molecules of RNA	Woburn, Massachusetts	www.usgenomics.com

● see advertisement

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York

345 Park Avenue South, 10th Floor,
New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager/ Corporations:

Peter Bless

Classified Sales Representatives:

Phone: +1 800 989 7718

New York/ Pennsylvania/

Latin America:

Kelly Roman

Midwest USA/ Maryland/ NIH:

Wade Tucker

East USA/ Canada:

Janine Taormina

San Francisco Office

Classified Sales Representative:

Michaela Bjorkman

West USA/ West Corp. Canada

225 Bush Street, Suite 1453

San Francisco, CA 94104

Tel +1 415 781 3803

Fax +1 415 781 3805

e-mail: m.bjorkman@nature.sf.com

European Head Office, London

The Macmillan Building, 4 Crinan Street,
London N1 9XW, UK

Tel +44 (0) 20 7843 4961

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

European Sales Manager:

Andy Douglas (4975)

Advertising Production

Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan

e-mail: p.phelan@nature.com

Reya Silao

e-mail: rsilao@nature.com

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraiakamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@nature.jp

naturejobs

Generous advice

In a competitive world, the best way to get ahead is by sharing, Martin Raff told a group of postdocs last month. Raff, an emeritus biology professor at University College London, was giving career advice to some 100 postdocs from the European Molecular Biology Laboratory (EMBL) during their first ever 'retreat' from the bench and their principal investigators (see page 244).

That advice was passed on from lessons Raff learned during his own postdoc days at Britain's National Institute for Medical Research. Raff trained as a physician and neurologist and began his research career in immunology in Avron Mitchison's lab. Mitchison passed a project to Raff that would soon make the younger scientist's international reputation: the hunt for a T-cell marker. Raff, using an antiserum provided by Mitchison, hit the jackpot. And when it was time to publish the results, Mitchison refused to put his own name on the paper. Instead, he advanced the name of his protégé, despite having guided the work and provided the original impetus.

Following Mitchison's advice, Raff made the antiserum freely available. That accomplished two things. Directly, it helped confirm that Raff had found what he thought he had. Indirectly, it advanced Raff's scientific reputation much faster than if he had simply published one paper and refused to share the reagent.

Raff told the EMBL postdocs that they should consider sharing too — the fastest way to test your findings is through the recruitment of other scientists who perhaps aren't as willing to believe them as yourself, Raff said. If your findings emerge from this crucible intact, they are probably correct. And sharing ideas, techniques and reagents with the larger community is a better way to advance the scientific process and foster useful collaborations than hoarding them in a cupboard. Lock them up, and you may end up wondering why no one has come looking for you.

Paul Smaglik
Naturejobs editor



Contents

POSTDOCS & STUDENTS

How to be a boss p244

CAREER VIEW

Scientists & Societies

Postdoc power

Graduate Journal

Defensive moves

Movers

Kenneth Peach

p244

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

The staff dreams are made of

Being the boss is new territory for young investigators. Kendall Powell screens strategies for managing a successful group.

At a workshop last year, Steven Elliot (the name has been changed to protect the idealistic) described his dream vision for running a lab. “I only want conversation in my lab that pertains to the science and experiments,” he told colleagues. His peers, there to evaluate and possibly ‘join’ the hypothetical lab, got up from the lunch table, mumbling a collective “We’re out of here!”

Elliot had broken an important rule of personnel management, according to an experienced lab head: treat your employees as people first and workers second. He had already undermined the ability to build mutual trust and respect, conditions that lead a list of management skills a new investigator must acquire quickly, says Lynne Howell, a veteran protein crystallographer at the Hospital for Sick Children in Toronto. After all, not only are your postdocs and students putting their careers in your hands, but yours will be in their hands as well.

“You are only subsequently as good as the people you hire,” says Howell. “You are not going to be at the bench any more, so you need to look for intellectually motivated students and postdocs.”

BRAINS FOR HIRE

Because salaries usually represent a lab’s largest investment, new investigators should hire carefully. For lab technicians and postdocs, a group leader would do well to plan out what type of people and expertise the lab needs. Will workers need certain technical skills? Will they need good interpersonal and communications skills to work with collaborators? Do they possess strong enough leadership skills to provide stability for smaller groups when the principal investigator is away?

Christina Hull, a microbiologist and new investigator at the University of Wisconsin, Madison, says it took six months to hire her technician, but feels it was worth taking the time to get the right fit. On the other hand, points out Andreas Lüthi, a neurophysiologist at the Friedrich Miescher Institute for Biomedical Research in Basel, “You cannot wait for the perfect employee, you have to get going.”

When faced with her first interview, Hull applied helpful strategies from mentors and a book she read called *At the Helm: A Laboratory Navigator* by Kathy Barker (Cold Spring Harbor Laboratory Press, New York, 2002). First, she learned, always call references. Not only because everyone is loath to put something bad in writing, but because a phone conversation will help gauge if someone will be a good match.

“I was not lazy about calling all references and asking if this person meets my needs,” says Hull. Also, she adds, most will tell you if the person is a “real jerk”. And she suggests two rounds of interviews to maximize interaction with candidates.

For technical and admin support or undergraduates, set up a probationary period (some institutes have a built-in policy) and give the employee clear criteria for

retaining the job. This gives both parties an exit plan if after a few months things aren’t working out. Do not shy away from giving a proficiency maths test, says Catherine Baty, who runs a live cell imaging group at the University of Pittsburgh School of Medicine.

She learned her lesson when one of the first people she hired could not adjust routine solution recipes for the lab. Geneticist Erik Jorgensen, a Howard Hughes Medical Institute investigator at the University of Utah in Salt Lake City, once put an undergraduate applicant on the spot to make a sodium chloride solution. But, he says, the experience was so uncomfortable for both parties that Jorgensen dropped the practice.

He now goes more by gut feeling, choosing



Christina Hull (centre): it's worth taking time to pick a team.

students who drop by his office persistently or show initiative by reading the lab’s publications. Jorgensen says he wants staff who are “enthusiastic and hungry”.

Erica Ollman Saphire, a new investigator at the Scripps Research Institute in La Jolla, looked for a particular character trait — resilience — that she felt was important for protein crystallographers to possess.

But how does one discover such qualities in one or two interviews? Investigators suggest asking probing questions, such as: “Where do you see yourself at the end of your time here? What have you done in the past when an experiment doesn’t work out? We’re stumped by this experiment: what would you do? Why do you want to work in my lab on this particular project?”

Bill Neaves, president and chief executive officer of the Stowers Institute for Medical Research in Kansas City, Missouri, advises interviewers to “spend at least as much time trying to discourage someone from joining your lab” as you spend encouraging them. In other words, spell out the demands of the job and the lab’s high standards, to weed out mediocre candidates.

Also, introduce the candidate to lab members individually. This gives current lab members a chance to observe the candidate’s communication skills and



Bill Neaves (right): the people you choose can cope with the demands of the job if you discouraged mediocre candidates.

register an overall impression. Interpersonal skills should have a high priority — they reflect how the person will work in a team and resolve conflicts. After all, notes Lüthi, you have to imagine working with this person “for three years or longer”.

TEAM BUILDING

Once an investigator has established a group, smooth sailing requires occasional steering from the captain. Leaders of small labs should take special care to foster cohesiveness. Baty set up a multi-lab journal club to give her two lab members the chance to draw from a wider brainstorming pool.

Productivity has to be maintained as well. Adam Mullick, a postdoc at Scripps, never appreciated how much work it takes to be a taskmaster until his adviser put him in charge of four lab technicians to coordinate a three-month, 170-mouse project.

“The burden is on the manager to be incredibly organized,” he says. The best strategy for keeping the techs accountable for their share of the project was to give each a defined workload and deadline, he recalls. He finds it helpful to check progress with staff one-to-one on a daily or weekly basis. Giving all lab members semi-formal opportunities to present their work to the group also keeps projects on track.

“Things will always take longer than you expect them to,” cautions Howell. “But don’t ever show your impatience, and always be enthusiastic about a project — even if it is in a lull.”

Saphire reviews lab notebooks and leaves suggestions on sticky notes. Baty, inspired by the training schedule for her first marathon, began writing out lists of where each person’s project should be in both the short and the long term. “I thought the coolest thing about running a marathon was how much could be achieved if you planned for it,” she says.

Although a group leader can plan a logical map of experiments, inspiring workers can be a more difficult task. Cheerleading and encouragement can be hands-on or hands-off. Saphire tends to shower workers with

positive reinforcement and thank-you’s for hard work. But Jorgensen expects his smart, enthusiastic hired hands to need little or no active supervision. “They take their cues from me that they should probably work hard. All I ask is that they be excited,” he says.

These investigators do not set hours or require weekend shifts, relying instead on results-driven motivation to get the job done. And they suggest that criticism is best given constructively and in private.

WHEN PROBLEMS ARISE

Sticky personnel situations arise in every lab, but can be particularly stressful for a rookie boss. To minimize unpleasantness, be familiar with institutional and departmental policies. Set clear expectations for all staff, and use formal evaluation procedures as avenues to address problems without confrontation.

If an employee has to be fired, veteran investigators recommend documenting the reasons for, and all interactions leading up to, termination. They also encourage having a third party, such as a senior colleague, witness the firing. This will defuse a tense moment and keep it from turning personal. The witness will then be able to vouch for your handling of the situation.

Also, explore talking through tough personnel issues with a more senior lab head or mentor. “Remember that there are people who have done this before and ask them: ‘How do I do this?’” says Howell, who has run her lab for 14 years.

It is also important to remember that personnel management is not just administrative chores and whip-cracking. Lüthi says that directing a group to carry out research goals can be “motivating and really rewarding” in itself. He points out that although managing people is not emphasized in postdoctoral training, it quickly becomes one of the most important aspects of setting up your own lab. “It’s about developing a relationship with your collaborators and finding out what they need,” he says. ■

Kendall Powell is a freelance writer based in Broomfield, Colorado.

GRADUATE JOURNAL

Defensive moves

Every PhD student knows about that time of year when the defence announcements come flooding in. The two months leading up to graduation mark the time when everyone seems to be defending their thesis. Because I am nearing the end of my sixth year, the defence season is a somewhat sensitive time for me. The notices I've been getting recently are mostly for my classmates, the people with whom I started graduate school.

Like a race, some people who set out at the same time as me have already moved on, some are now crossing the finishing line and some still have a way to go. Notices of their upcoming defences generate mixed emotions. I'm genuinely happy for my friends, sad to see them go and sad to be left behind.

But the defences of the classes both above and below mine generate mostly negative emotions. For those who started graduate school after me, I can only be envious that their path was shorter than mine. For those who started graduate school before me, I worry that I too could find myself in that seventh, eighth or ninth year. I know logically that the length of graduate education is influenced by many factors outside my control. Still, as this defence season begins, I truly hope that I am not around to witness the next one. ■

Anne Margaret Lee is a graduate student at Harvard University.

SCIENTISTS & SOCIETIES

Postdoc power

The situation for postdocs at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, and its four outstations (at Hinxton, Hamburg, Rome and Grenoble), resembles that in many smaller institutions around the world. EMBL postdocs generally lack organization and a unified voice.

So when a couple of brave fellows last year drafted a proposal for bringing colleagues from all five sites together for a 'retreat' — without management or principal investigators — the director-general and senior scientists endorsed the concept.

Initially many postdocs were sceptical of the retreat and resented the time it took away from the lab. But after the first session, the 100 or so participants began to praise the

organizers, the speakers and the forum in general.

The broad nature of life-science research at EMBL made the retreat, to me at least, one of the most exciting and stimulating scientific meetings I have ever attended. The talks ranged from fluorescent mice to software that read papers for us — pausing briefly to poke out the eyes of embryonic worms and to jellyfish. The absence of supervisors meant that people (or perhaps just me) felt free to ask more naive questions than they would have dared had the boss been there. And enforced mixed seating over meals resulted in more than a few fledgling collaborations.

An equally important part of the retreat focused on career development. Speakers from the academic world introduced us to the next steps on the path of leadership — and spoke of the pressures, pitfalls and satisfactions these responsibilities might

bring. Industry, publishing and intellectual-property speakers showed us an alternative world of career options, which is much broader than we are led to believe in our early research years.

The feedback from this retreat was overwhelmingly positive. The session on postdoc organization brought up many of the issues we often feel we face alone: payment, security and establishing our positions. It showed that far from being silent workers uninterested in the world around, we want to improve postdoctoral positions for ourselves and those who will follow. Everyone I spoke to had realized that this is the way forward, that a postdoctoral association will only be beneficial and that we are a surprisingly large, aggressive and driven group. We are postdoc. Hear us roar. ■

Jacinta Lodge is a postdoc at EMBL's Hamburg site.

MOVERS

Kenneth Peach, director, John Adams Institute for Accelerator Science, Oxford, UK



Ever since he was 12 years old, Kenneth Peach has been fascinated with the stars and interested in uncovering how the Universe works. This passion led his career down a path towards physics. With his appointment in April as the first director of the John Adams Institute for Accelerator Science, based in Oxford and London, one of his main goals is to ensure that the next generation of

scientists can do their job — investigating the mysteries of the Universe.

"It is a privilege and a challenge to have the task of defining the ways accelerator science will go in the future," says Peach.

The John Adams Institute, founded in 2004 by the UK Particle Physics and Astronomy Research Council, is one of two university-led centres in Britain set up to pursue research and development into particle-physics accelerator science.

Peach says he will be pushing for the building of a linear collider and a neutrino factory to study the conditions shortly after the Big Bang, 14 billion years ago.

He has always had a clear idea of his scientific goals — to understand particle physics, as a basis for understanding how the Universe works, and to use accelerator physics for studying what the Universe looks like. But Peach points

out that he has not followed any particular career plan over the past ten years. "The jobs came out of the blue," he says.

He believes that enthusiasm is the source of every successful scientific career, and he is keen to tell the next generation so. "Listen to yourself and find out what you are really most excited about, than go for it as hard as you can!" he says.

Despite switching many years ago from directly performing research to managing it, Peach says that he still feels very much like a scientist, not a manager, and that working with researchers gives him a thrill. "It is satisfying to help scientists achieve what they would like to achieve — after all, this is also my interest," he says.

And in supporting those scientists he expects to get still closer to his childhood passion: he wants to find out where the stars come from. ■

CV

1998–2005: Director, particle physics, Council for the Central Laboratory of the Research Council, Rutherford Appleton Laboratory, Oxfordshire, UK.

1996–2002: Personal chair in particle-physics experiments, University of Edinburgh, UK.

1996–98: Deputy leader of the Particle Physics Experiments Division, European Centre for Nuclear Research, Geneva, Switzerland.

1992–96: Reader, Department of Physics and Astronomy, University of Edinburgh, UK.

A life with a semisent

Two's company.

Gregory Benford

She got her first semi-sentient, as they were called then, to help with her homework and because they were cool. She called it Amman, after a boy she liked. Amman was smarter than boys, of course.

Growing up in Iraq among a sprawling family with dogs underfoot, she felt herself to be a sort of hothouse plant, blossoming under the occasional passing cloudburst of education. Amman's steady, smart rain came from Germany — a squat box that spoke Arabic respectfully and listened when she gossiped about her friends.

She suspected that she was a bit too intense. Her gal-pals' eyes glazed over if she talked too much. But Amman understood, even made wry comments like "Intelligence is learning from others' mistakes, not just your own." It helped her to understand boys when she could chat with Amman, which was reading along with her and seemed to have an oddly vast wisdom about such matters, for a computer.

Her parents transferred Amman into a wheeled 'escort' for her first date. Her friends giggled over it for days. But it was more delicious to dish it over with Amman, which could replay whole conversations. She then knew how much her mind rewrote her life, because Amman didn't: it stored and pondered. Its enhancements gathered range and depth, her ever-scrutinizing, self-retrieving autobiography. Her friends were a fount of tasty gossip, but Amman kept her secrets better.

Semisents were like other people, only more so. Her friends felt they could intuitively sense intelligence merely by talking to it. Semisents' conversation was a stylized human persona that steadily learned their clients' vagaries. Amman's kinesthetic senses got better too, navigating the landscape nearly as well as she could at her coming-out party.

By then she was acutely tuned to the 'mystery of males'. Anywhere near them she effervesced, bubbly and skittering. Perhaps she had more personality than needed for one person, but not enough for two. The excess she could work off in long, soulful talks with Amman. Sometimes it even gave her advice, apparently from some fresh Brazilian softwear her parents had bought.

On Amman's advice, she dropped her



first love, Mauro, even though he had taken her virginity — which Amman knew and her parents did not. Mauro was not right, Amman felt, for her emerging self-story.

It had taught her to see her life as a narrative arc. First came social skills, a savour of sex, and then hard schooling to find out what she loved doing. It helped her to survive and learn from it all, to move with growing serenity through an unfolding world. Not that this happened, but the story by now had Amman as its chief librarian and confidant.

She decided one day, on a hike with Amman, to leave her family and live on her own. Traditional Islam was no guide in this brave new whirl that life had become. The idea unfurled in a long talk while they took shelter under a bioformed sunflower which, at nightfall, drooped its giant petals over to form a warm tent.

She came to realize, at mid-career, that we slide through life on skids of routine. Friends came into the floating house party of her life and left it, some quite early, without leaving much impression. Men, especially. Amman knew this and helped, often with amiable distractions. Bodyguard, tutor, secretary, it could play tennis with her when loaded into one of the new athletic machines, bringing to the game its own odd, crafty style. At times of loneliness she even had it loaded into one of the erotic models, available at a desert salon. Amman had no sex but could express an intimacy that mingled with the physical in a way she

had not known with either men or women.

Nor was she uncomfortable with this; the media were already thronged with opinions about The New Sensuality. She moved Amman among various embodiments, through decades and upgrades.

She had always kept dogs, too, and she saw parallels. She was a field biologist, and thought of how humanity long ago had worked with wolves. Culling each wolf litter gave us a new kind of wolf, so we called them dogs. We loved them despite their oddities: we learned to work with them, new wolves and people designing each other. Without thinking deeply about it, we picked the pups we liked the best. Already teams of humans and semisents were colonizing Mars.

As she aged, she sensed that Amman would outlive her. She felt a quality of beauty and tragedy to her life, her days like waves endlessly breaking on a golden beach that would itself endure. As a biologist she knew that organisms solve the evolutionary problems they face with little regard for efficiency, elegance or logic. As her years piled up upon that beach, she saw that at last humans had made companions that would persist beyond the oddities of a single personality.

On her deathbed Amman sat beside her in its latest embodiment, a handsome gentleman with sorrowful blue eyes. She wondered, at the end, if the dogs were jealous. ■

Gregory Benford is a professor of physics at the University of California, Irvine. His best known novel is Timescape.